

# A Bayesian model selection approach for identifying differentially expressed transcripts from RNA-Seq data

Panagiotis Papastamoulis

panagiotis.papastamoulis@manchester.ac.uk

*University of Manchester, Faculty of Life Science, Manchester, UK*

Magnus Rattray

magnus.rattray@manchester.ac.uk

*University of Manchester, Faculty of Life Science, Manchester, UK*

**Summary.** Recent advances in molecular biology allow the quantification of the transcriptome and scoring transcripts as differentially or equally expressed between two biological conditions. Although these two tasks are closely linked, the available inference methods treat them separately: a primary model is used to estimate expression and its output is post-processed using a differential expression model. In this paper, both issues are simultaneously addressed by proposing the joint estimation of expression levels and differential expression: the unknown relative abundance of each transcript can either be equal or not between two conditions. A hierarchical Bayesian model builds upon the BitSeq framework and the posterior distribution of transcript expression and differential expression is inferred using Markov Chain Monte Carlo (MCMC). It is shown that the proposed model enjoys conjugacy for fixed dimension variables, thus the full conditional distributions are analytically derived. Two samplers are constructed, a reversible jump MCMC sampler and a collapsed Gibbs sampler, and the latter is found to perform best. A cluster representation of the aligned reads to the transcriptome is introduced, allowing parallel estimation of the marginal posterior distribution of subsets of transcripts under reasonable computing time. The proposed algorithm is benchmarked against alternative methods using synthetic datasets and applied to real RNA-sequencing data. Source code is available online <sup>†</sup>.

**Keywords:** RNA-sequencing, mixture models, collapsed Gibbs, reversible jump MCMC

## 1. Introduction

Quantifying the transcriptome of a given organism or cell is a fundamental task in molecular biology. RNA-sequencing (RNA-Seq) technology produces transcriptomic data in the

<sup>†</sup><https://github.com/mqbssppe/cjBitSeq>

form of short reads (Mortazavi et al., 2008). These reads can be used either in order to reconstruct the transcriptome using *de novo* or guided assembly, or to estimate the abundance of known transcripts given a reference annotation. Here, we consider the latter scenario in which transcripts are defined by annotation. In such a case, millions of short reads are aligned to the reference transcriptome (or genome) using mapping tools such as Bowtie (Langmead et al., 2009) (or Tophat (Trapnell et al., 2009)). Of particular interest is the identification of differentially expressed transcripts (or isoforms) across different samples. Throughout this paper the term transcript refers to isoforms, so differential transcript detection has the same meaning as differential isoform detection. Most genes in higher eukaryotes can be spliced into alternative transcripts that share specific parts of their nucleotide sequence. Thus, a short read is not uniquely aligned to the transcriptome and its origin remains uncertain, making transcript expression estimation non-trivial. Probabilistic models provide a powerful means to estimate transcript abundances as they are able to take this ambiguous read assignment into consideration in a principled manner.

There are numerous methods that estimate transcript expression from RNA-Seq data, including RSEM (Li and Dewey, 2011), IsoEM (Nicolae et al., 2011), Cufflinks (Trapnell et al., 2010, 2013), BitSeq (Stage 1)(Glaus et al., 2012), TIGAR (Nariai et al., 2013) and Casper (Rossell et al., 2014). Some of these methods also include a second stage for performing DE analysis at the transcript level (e.g. Cuffdiff and BitSeq Stage 2) and stand-alone methods for transcript-level DE calling have also been developed such as EBSeq (Leng et al., 2013) and MetaDiff (Jia et al., 2015). Cuffdiff uses an asymptotically normal test statistic by applying the delta method to the log-ratio of transcript abundances between two samples, given the estimated expression levels using Cufflinks. EBSeq estimates the Bayes factor of a model under DE or nonDE for each transcript, building a Negative Binomial model upon the estimated read counts from any method. BitSeq Stage 2 ranks transcripts as differentially expressed by the probability of positive log-ratio (PPLR) based on the MCMC output from BitSeq Stage 1, which estimates the expression levels assuming a mixture model. Gene-level DE analysis is also available using count-based methods such as edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010) but here we limit our attention to methods designed for transcript-level DE calling.

All existing methods for transcript-level DE calling apply a two-step procedure. The mapped RNA-Seq data is used as input of a first stage analysis to estimate transcript expression. The output of this stage is then post-processed at a second stage in order to classify transcripts as DE or non-DE. The bridge between the two stages is based

upon certain parametric assumptions for the distribution of the estimates of the first stage and/or the use of asymptotic results (as previously described above). Also, transcript-level expression estimates are correlated through sharing of reads and this correlation is typically ignored in the second stage. Such two-stage approaches are quite useful in practice since the differential expression question is not always the main aim of the analysis; therefore estimating expression is useful in itself. However, when the main purpose of an experiment is DE calling then the two-stage procedure increases the modelling complexity and may result in overfitting, since there is no guarantee that the underlying assumptions are valid. Note that a recent method (Gu et al., 2014) addresses the joint estimation of expression and differential expression modelling of exon counts under a Bayesian approach but at the gene level rather than the transcript level considered here.

The contribution of this paper is to develop a method for the joint estimation of expression and differential expression at the transcript level. The method builds upon the Bayesian framework of the BitSeq (Stage 1) model where transcript expression estimation reduces to estimating the posterior distribution of the weights of a mixture model using MCMC Glaus et al. (2012). The novelty in the present study is that differential expression is addressed by inferring which weights differ between two mixture models. This is achieved by using two samplers. A reversible jump MCMC (rjMCMC) algorithm (Green, 1995) updates both transcript expression and differential expression parameters, while a collapsed Gibbs algorithm is developed which avoids transdimensional transitions. The high-dimensional setting of RNA-seq data studies makes the convergence to the joint posterior distribution computationally challenging. To alleviate this computational burden and allow easier parallelization, a new cluster representation of the transcriptome is introduced which collapses the problem to subsets of transcripts sharing aligned reads.

The rest of the paper is organized as follows. The mixture model used in the original BitSeq setup is reviewed in Section 2.1. The prior assumptions of the new cjBitSeq (clusterwise joint BitSeq) model is introduced in Section 2.2. The full conditional distributions are given in Section 2.3 and two MCMC samplers are described in Section 2.4. A cluster representation of aligned reads and transcripts is discussed in Section 2.5 and details over False Discovery Rate (FDR) estimation are given in Section 2.6. Large scale simulation studies are presented in Section 3.2 and the proposed method is illustrated to a real human dataset in Section 3.3. The paper concludes in Section 4 with a synopsis and discussion.

## 2. Methods

In the BitSeq model, the mixture components correspond to annotated transcript sequences and the mixture weights correspond to their relative expression levels. The data likelihood is then computed by considering the alignment of reads (or read-pairs) against each mixture component. Essentially, this model is modified here in order to construct a well-defined probability of DE or non-DE when two samples are available.

We induce a set of free parameters of varying dimension, depending on the number of different weights between two mixture models. Assuming two independent Dirichlet prior distributions, the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) draws samples from the full conditionals, which are independent Dirichlet and Generalized Dirichlet (Connor and Mosimann, 1969; Wong, 1998, 2010) distributions. This representation allows the integration of the corresponding parameters as stated at Theorem 2. Therefore, we provide two MCMC samplers depending on whether transcript expression levels are integrated out or not. These samplers converge to the same target distribution but using different steps in order to update the state of each transcript: the first one uses a birth-death move type (Richardson and Green, 1997; Papastamoulis and Iliopoulos, 2009) and the second one is a block update from the full conditional distribution. After detecting clusters of transcripts and reads, it is shown that the parallel application of the algorithm to each cluster converges to proper marginals of the full posterior distribution.

### 2.1. BitSeq

Let  $\mathbf{x} = (x_1, \dots, x_r)$ ,  $x_i \in \mathcal{X}$ ,  $i = 1, \dots, r$ , denote a sample of  $r$  short reads aligned to a given set of  $K$  transcripts. The sample space  $\mathcal{X}$  consists of all sequences of letters A, C, G, T. Assuming that reads are independent, the joint probability density function of the data is written as

$$\mathbf{x}|\boldsymbol{\theta} \sim \prod_{i=1}^r \sum_{k=1}^K \theta_k f_k(x_i). \quad (1)$$

The number of components ( $K$ ) is equal to the number of transcripts and it is considered as known since the transcriptome is given. The parameter vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \in \mathcal{P}_{K-1}$  denotes relative abundances, where

$$\mathcal{P}_{K-1} := \{p_k \geq 0, k = 1, \dots, K-1 : \sum_{k=1}^{K-1} p_k \leq 1; p_K := 1 - \sum_{k=1}^{K-1} p_k\}.$$

The component specific density  $f_k(\cdot)$  corresponds to the probability of a read aligning at some position of transcript  $k$ ,  $k = 1, \dots, K$ . Since we assume a known transcriptome,

$\{f_k\}_{k=1}^K$  are known as well and they are computed according to the methodology described in Glaus et al. (2012) (see also Appendix A in supplementary material), taking into account position and sequence-specific bias correction methods.

A priori it is assumed that  $\boldsymbol{\theta} \sim \mathcal{D}_{K-1}(\alpha_1, \dots, \alpha_K)$ , with  $\mathcal{D}_j$  denoting the Dirichlet distribution defined over  $\mathcal{P}_j$ . Furthermore, it is assumed that  $\alpha_1 = \dots = \alpha_K = 1$ , which is equivalent to the uniform distribution in  $\mathcal{P}_{K-1}$ . In the original implementation of BitSeq (Glaus et al., 2012), MCMC samples are drawn from the posterior distribution of  $\boldsymbol{\theta}|\mathbf{x}$  using the Gibbs sampler while more recently variational Bayes approximations have also been included for faster inference (Papastamoulis et al., 2014; Hensman et al., 2015).

Given the output of BitSeq stage 1 for two different samples, BitSeq stage 2 implements a one-sided test (PPLR) for DE analysis. However, this approach does not define transcripts as DE or non-DE and is therefore not directly comparable to standard 2-sided tests available in most other packages (Trapnell et al., 2013; Leng et al., 2013). Also, correlations between transcripts in the posterior distribution for each sample are discarded during the DE stage, leading to potential loss of accuracy when making inferences. In order to deal with these limitations, a new method for performing DE analysis is presented next.

## 2.2. cjBitSeq

Assume that we have at hand two samples  $\mathbf{x} := (x_1, \dots, x_r)$  and  $\mathbf{y} := (y_1, \dots, y_s)$  denoting the number of (mapped) reads for sample  $\mathbf{x}$  and  $\mathbf{y}$ , respectively. Now, let  $\theta_k$  and  $w_k$  denote the unknown relative abundance of transcript  $k = 1, \dots, K$  in sample  $\mathbf{x}$  and  $\mathbf{y}$ , respectively. Define the parameter vector of relative abundances as  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{K-1}; \theta_K) \in \mathcal{P}_{K-1}$  and  $\mathbf{w} = (w_1, \dots, w_{K-1}; w_K) \in \mathcal{P}_{K-1}$ . Under the standard BitSeq model the prior on the parameters  $\boldsymbol{\theta}$  and  $\mathbf{w}$  would be a product of independent Dirichlet distributions. In this case the probability  $\theta_k = w_k$  under the prior is zero and it is not straightforward to define non-DE transcripts. To model differential expression we would instead like to identify instances where transcript expression has not changed between samples. Therefore, we introduce a non-zero probability for the event  $\theta_k = w_k$ . This leads us to define a new model with a non-independent prior for the parameters  $\boldsymbol{\theta}$  and  $\mathbf{w}$ .

**DEFINITION 1 (STATE VECTOR).** Let  $\mathbf{c} := (c_1, \dots, c_K) \in \mathcal{C}$ , where  $\mathcal{C}$  is the set defined by:

- (a)  $c_k \in \{0, 1\}$ ,  $k = 1, \dots, K$
- (b)  $c_+ := \sum_{k=1}^K c_k \neq 1$ .

Then, for  $k = 1, \dots, K$  let: 
$$\begin{cases} \theta_k = w_k, & \text{if } c_k = 0 \\ \theta_k \neq w_k, & \text{if } c_k = 1. \end{cases}$$

We will refer to vector  $c$  as the state vector of the model.

For example, assume that  $K = 6$  and  $c = (1, 0, 0, 1, 0, 1)$ . According to Definition 1,  $\theta_k = w_k$  for  $k = 2, 3, 5$  and  $\theta_k \neq w_k$  for  $k = 1, 4, 6$ . From Definition 1 it is obvious that the sum of the elements in  $c$  cannot be equal to 1 because either all  $\theta$ 's have to be equal to  $w$ 's, or at least two of them have to be different. The introduction of such dependencies between the elements of  $\theta$  and  $w$  has non-trivial effects on the prior assumptions of course. It is clear that with this approach we should define a valid conditional prior distribution for  $\theta, w|c$ .

At first we impose a prior assumption on  $c$ . We will consider the Jeffreys' (Jeffreys, 1946) prior distribution for a Bernoulli trial, that is  $P(c_k = 1|\pi) = \pi$  with  $\pi$  following a Beta distribution. Since  $c_+ \neq 1$ , the prior distribution of the state vector  $c$  is expressed as

$$\pi \sim \text{Beta}(1/2, 1/2) \quad (2)$$

$$P(c|\pi) = P(c|c_+ \neq 1, \pi) = \frac{\pi^{c_+}(1-\pi)^{K-c_+}}{1 - K\pi(1-\pi)^{K-1}}, \quad c \in \mathcal{C}. \quad (3)$$

Next we proceed to the definition of a proper prior structure for the weights of the mixture. At this step extra care should be taken for everything to make sense as a probabilistic space. It is obvious that  $(\theta, w)$  should be defined conditional to the state vector  $c$ . What it is less obvious, is that  $(\theta, w)$  should be defined conditional on a parameter of varying dimension. At this point, we introduce some extra notation.

**DEFINITION 2 (DEAD AND ALIVE SUBSETS AND PERMUTATION OF THE LABELS).** *For a given state vector  $c$ , define the order-specific subsets*

$$C_0(c) := \{\tau_1 < \dots < \tau_{K-c_+} \in \{1, \dots, K\} : c_{\tau_k} = 0 \quad \forall k = 1, \dots, K - c_+\}$$

and

$$C_1(c) := \{\tau_{K-c_++1} < \dots < \tau_K \in \{1, \dots, K\} : c_{\tau_k} = 1 \quad \forall k = K - c_+ + 1, \dots, K\}.$$

*These sets will be called dead and alive subsets of the transcriptome index, respectively. Moreover,  $\tau = (\tau_1, \dots, \tau_K)$  denotes the unique permutation of  $\{1, \dots, K\}$  obeying the ordering within the dead and alive subsets.*

As it will be made clear later, it is convenient to define a unique labelling within the dead and alive subsets so we also explicitly defined the corresponding permutation ( $\tau$ ) of the

labels. In order to clarify Definition 2, assume that  $c = (1, 0, 0, 1, 0, 1)$ . Then Definition 2 implies that  $C_0(c) = \{2, 3, 5\}$ ,  $C_1(c) = \{1, 4, 6\}$  and  $\tau = (2, 3, 5, 1, 4, 6)$ . The order-specific definition of these subsets excludes  $\{3, 2, 5\}$  (for example) from the definition of a dead subset.

It is clear that if  $C_0(c) = \emptyset$ , then both  $\boldsymbol{\theta}$  and  $\boldsymbol{w}$  have  $K - 1$  free parameters each. However, if  $C_0(c) \neq \emptyset$ , the free parameters are lying in a lower dimensional space. This means that  $(\boldsymbol{\theta}, \boldsymbol{w})$  should be defined given  $c$  by taking into account the set of free parameters that are actually allowed by the state vector. In particular,  $(\boldsymbol{\theta}, \boldsymbol{w})$  are pseudo-parameters. The actual parameters of our problem are defined in Lemma 1.

In what follows, the notation  $\tau\boldsymbol{\sigma}$  should be interpreted as the reordering of vector  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K)$  under permutation  $\tau$ . E.g: assume that  $\tau = (3, 1, 2)$  and  $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ , then:  $\tau\boldsymbol{\sigma} = (\sigma_3, \sigma_1, \sigma_2)$ . Let also  $\tau^{-1}$  denote the inverse permutation of  $\tau$ .

LEMMA 1 (EXISTENCE AND UNIQUENESS OF FREE PARAMETERS). *For every  $(c, \tau, \boldsymbol{\theta}, \boldsymbol{w})$  respecting Definitions 1 and 2 there exists a unique set of free parameters:*

$$(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{P}_{K-1} \times \mathcal{P}_{c_+-1}, \quad (4)$$

such that:

$$\boldsymbol{\theta} = \tau^{-1}\boldsymbol{u} \quad (5)$$

$$\boldsymbol{w} = \tau^{-1}\boldsymbol{\varpi}, \quad (6)$$

where  $\boldsymbol{\varpi} = \left( \{u_{\tau_k^{-1}} : k \in C_0(c)\}, \boldsymbol{v} \sum_{k \in C_1(c)} u_{\tau_k^{-1}} \right)$  under the conventions  $\mathcal{P}_{-1} := \emptyset$  and  $\emptyset \sum_{k \in \emptyset} u_k := \emptyset$ .

PROOF. It is trivial to show that  $(c, \tau, \boldsymbol{u}, \boldsymbol{v}) \rightarrow (\boldsymbol{\theta}, \boldsymbol{w})$  is an “one to one” and “onto” mapping (bijective function).

**Example:** Assume that  $c = (1, 0, 0, 1, 0, 1)$ , where  $C_0(c) = \{2, 3, 5\}$  and  $C_1(c) = \{1, 4, 6\}$ . Then,  $\tau = (2, 3, 5, 1, 4, 6)$  and  $\tau^{-1} = (4, 1, 2, 5, 3, 6)$ . According to state  $c$  we should have that  $\theta_2 = w_2$ ,  $\theta_3 = w_3$  and  $\theta_5 = w_5$ , while  $\theta_k \neq w_k$  for  $k \in C_1(c)$ . Lemma 1 states that  $\boldsymbol{\theta}$  and  $\boldsymbol{w}$  can be expressed as a transformation of two independent parameters:  $\boldsymbol{u} = (u_1, u_2, u_3, u_4, u_5, u_6) \in \mathcal{P}_5$  and  $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathcal{P}_2$ . According to Equation (5),  $\boldsymbol{\theta}$  is a permutation of the vector  $\boldsymbol{u}$ :

$$\boldsymbol{\theta}|(c, \boldsymbol{u}) = (u_4, u_1, u_2, u_5, u_3, u_6).$$

Next,  $\boldsymbol{w}$  is obtained by a permutation of  $\boldsymbol{\varpi}$ , which is a linear transformation of  $\boldsymbol{u}$  and  $\boldsymbol{v}$ , that is,  $\boldsymbol{\varpi} = (u_1, u_2, u_3, v_1(u_4 + u_5 + u_6), v_2(u_4 + u_5 + u_6), v_3(u_4 + u_5 + u_6))$ . According

to Equation (6):

$$\mathbf{w}|(c, \mathbf{u}, \mathbf{v}) = (v_1(u_4 + u_5 + u_6), u_1, u_2, v_2(u_4 + u_5 + u_6), u_3, v_3(u_4 + u_5 + u_6)).$$

Comparing the last two expressions for  $\boldsymbol{\theta}$  and  $\mathbf{w}$ , it is obvious that  $\theta_2 = w_2$ ,  $\theta_3 = w_3$  and  $\theta_5 = w_5$ , while  $\theta_k \neq w_k$  for all remaining entries, which is the configuration implied by the state vector  $c$ . Note also that  $\{u_{\tau_k^{-1}}; k \in C_0(c)\} = (u_1, \dots, u_{K-c_+})$  and  $\{u_{\tau_k^{-1}}; k \in C_1(c)\} = (u_{K-c_++1}, \dots, u_K)$  and  $\sum_{k \in C_1(c)} w_k = \sum_{k \in C_1(c)} \theta_k = \sum_{k \in C_1(c)} u_{\tau_k^{-1}}$ .

Now, it should be clear that given a state vector  $c$ , as well as the independent free parameters  $\mathbf{u}$  and  $\mathbf{v}$ , the pseudo-parameters  $\boldsymbol{\theta}$  and  $\mathbf{w}$  are deterministically defined. In other words, the conditional distributions of  $\boldsymbol{\theta}$  and  $\mathbf{w}$  are Dirac, gathering all their probability mass into the single points defined by Equations (5) and (6). Hence, the conditional prior distribution for transcript expression is written as:

$$f(\boldsymbol{\theta}, \mathbf{w}|c, \tau, \mathbf{u}, \mathbf{v}) = 1_{\boldsymbol{\theta}, \mathbf{w}}(\{\boldsymbol{\theta}(c, \tau, \mathbf{u}), \mathbf{w}(c, \tau, \mathbf{u}, \mathbf{v})\}), \quad (7)$$

with  $\boldsymbol{\theta}(c, \tau, \mathbf{u})$  and  $\mathbf{w}(c, \tau, \mathbf{u}, \mathbf{v})$  as in Equations (5) and (6), respectively.

Moreover, we stress that if the permutation  $\tau$  was not uniquely defined according to Definition 2, then we would have had to take into account all the possible permutations within the dead and alive subsets. However, such an approach would lead to an increased modelling complexity without making any difference at the inference. That said, the conditional prior distribution of  $\tau$  given  $c$  is Dirac:

$$f(\tau|c) = 1_{\tau}(\tau(c)), \quad (8)$$

where  $\tau(c)$  denotes the unique permutation (given  $c$ ) in Definition 2.

At this point we state our prior assumptions for the free parameters, given a state vector  $c$ . We assume that a priori  $\mathbf{u}$  and  $\mathbf{v}$  are independent random variables distributed according to Dirichlet distribution, that is:

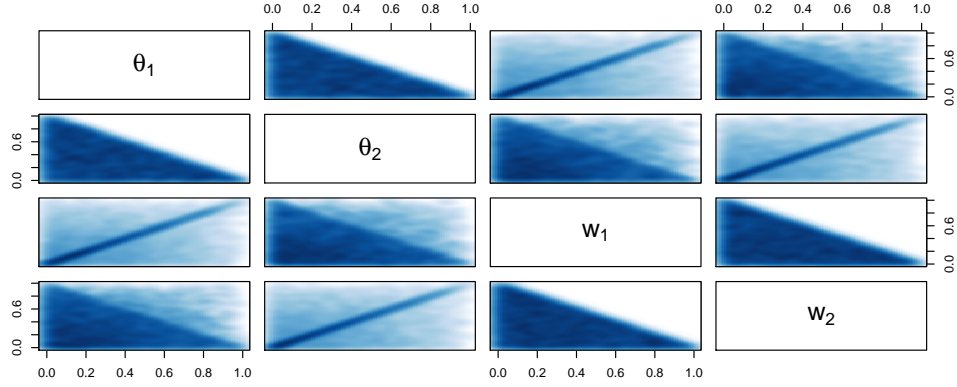
$$\mathbf{u}|c \sim \mathcal{D}_{K-1}(\alpha_1, \dots, \alpha_K) \quad (9)$$

$$\mathbf{v}|c \sim \mathcal{D}_{c_+-1}(\gamma_1, \dots, \gamma_{c_+}). \quad (10)$$

In the applications, we will furthermore assume that  $\alpha_k = 1$  for all  $k = 1, \dots, K$  and  $\gamma_\ell = 1$  for all  $\ell = 1, \dots, c_+$ , in order to assign a uniform prior distributions over  $\mathcal{P}_{K-1} \times \mathcal{P}_{c_+-1}$ . Now, the following Theorem holds.

**THEOREM 1.** *Assume that (9) and (10) hold true and furthermore:  $\alpha_k = \gamma_k = \alpha$  for all  $k = 1, \dots, K$ . Then,  $\boldsymbol{\theta}$  and  $\mathbf{w}$  are marginally identical random variables following the  $\mathcal{D}_{K-1}(\alpha, \dots, \alpha)$  distribution.*





**Fig. 1.** Simulation from the prior distribution (7) of  $(\boldsymbol{\theta}, \boldsymbol{w})$  for  $K = 3$ ,  $\alpha_k = \gamma_k = 1$  for  $k = 1, 2, 3$ , and also assuming the Jeffreys' prior for  $c$ . Theorem 1 states that marginally:  $\boldsymbol{\theta} \sim \mathcal{D}(1, 1, 1)$  and  $\boldsymbol{w} \sim \mathcal{D}(1, 1, 1)$ .

PROOF. See Appendix C in supplementary material.

Note here that Theorem 1 does not imply that  $\boldsymbol{\theta}$  and  $\boldsymbol{w}$  are a priori independent. As shown in Figure 1,  $\theta_k$  is exactly equal to  $w_k$  with probability  $P(c_k = 0) > 0$ ,  $k = 1, \dots, K$ .

The model definition is completed by considering the latent allocation variables of the mixture model. Let  $\boldsymbol{\xi} = \{\xi_1, \dots, \xi_r\}$  and  $\boldsymbol{z} = \{z_1, \dots, z_s\}$  with

$$\begin{aligned} P(\xi_i = k | \boldsymbol{\theta}) &= \theta_k, \quad \text{independent for } i = 1, \dots, r \\ P(z_j = k | \boldsymbol{w}) &= w_k, \quad \text{independent for } j = 1, \dots, s, \end{aligned}$$

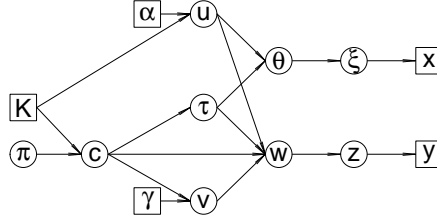
for  $k = 1, \dots, K$ . Moreover,  $\boldsymbol{\xi}, \boldsymbol{z}$  are assumed conditionally independent given  $\boldsymbol{\theta}$  and  $\boldsymbol{w}$ , that is,  $P(\boldsymbol{\xi}, \boldsymbol{z} | \boldsymbol{\theta}, \boldsymbol{w}) = P(\boldsymbol{\xi} | \boldsymbol{\theta})P(\boldsymbol{z} | \boldsymbol{w})$ . Now, the joint distribution of the complete data  $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z})$  factorizes as follows:

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z} | \boldsymbol{\theta}, \boldsymbol{w}) = \prod_{i=1}^r \theta_{\xi_i} f_{\xi_i}(x_i) \prod_{j=1}^s w_{z_j} f_{z_j}(y_j). \quad (11)$$

Let  $\boldsymbol{g} = (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}, c, \tau, \pi)$ . From Equations (2), (3) and (7)-(11), the joint distribution of  $\boldsymbol{g}$  is defined as

$$\begin{aligned} f(\boldsymbol{g} | \boldsymbol{\alpha}, \boldsymbol{\gamma}, K) &= f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z} | \boldsymbol{\theta}, \boldsymbol{w}) f(\boldsymbol{u} | \boldsymbol{\alpha}, K) f(\boldsymbol{v} | c, \boldsymbol{\gamma}) f(\boldsymbol{\theta} | \tau, \boldsymbol{u}) \\ &\times f(\boldsymbol{w} | c, \tau, \boldsymbol{u}, \boldsymbol{v}) f(\tau | c) f(c | K, \pi) f(\pi). \end{aligned} \quad (12)$$

Equation (12) defines a hierarchical model whose graphical representation is given in Figure 2 with circles (squares) denoting unobserved (observed/known) variables.



**Fig. 2.** Directed Acyclic Graph representation of the hierarchical model (12).

### 2.3. Full conditional distributions for the Gibbs updates

In this section, the full conditional distributions are derived. Let  $h|\cdots$  denote the conditional distribution of a random variable  $h$  given the values of the rest of the variables. We also denote by  $\mathbf{x}_{[-i]}$  all remaining members of a generic vector after excluding its  $i$ -th item.

It is straightforward to show that  $\pi|\cdots \sim \text{Beta}(c_+ + 1/2, K - c_+ + 1/2)$ . For the allocation variables it follows that:

$$P(\xi_i = k|\cdots) \propto \theta_k f_k(x_i) \quad k = 1, \dots, K \quad (13)$$

$$P(z_j = k|\cdots) \propto w_k f_k(y_j) \quad k = 1, \dots, K \quad (14)$$

independent for  $i = 1, \dots, r$  and  $j = 1, \dots, s$ . Now, given  $(\mathbf{u}, \mathbf{v}, c, \tau)$ , it is again trivial to see that the full conditional distributions of  $\boldsymbol{\theta}, \mathbf{w}|\cdots$  is the same as in (7). Let  $\mathcal{GD}(\cdot, \cdot)$  denotes the Generalized Dirichlet distribution (see Appendix B in supplementary material) and also define

$$s_k(\boldsymbol{\xi}) := \sum_{i=1}^r I(\xi_i = k), \quad s_k(\mathbf{z}) := \sum_{j=1}^s I(z_j = k)$$

for  $k = 1, \dots, K$ . Regarding the full conditional distribution of the free parameters, we have the following result.

LEMMA 2. *The full conditional distribution of  $(\mathbf{u}, \mathbf{v}|\cdots)$  is*

$$\mathbf{u}|\cdots \sim \mathcal{GD}(\lambda_1, \dots, \lambda_{K-1}; \beta_1, \dots, \beta_{K-1}) \quad (15)$$

$$\mathbf{v}|\cdots \sim \mathcal{D}_{c_+-1}(\{\gamma_\ell + s_{\tau_{\ell+k^*}}(\mathbf{z}); \ell = 1, \dots, c_+\}), \quad (16)$$

with  $k^* := K - c_+$ , conditionally independent (given all other variables), where

$$\lambda_k := \begin{cases} \alpha_k + s_{\tau_k}(\boldsymbol{\xi}) + s_{\tau_k}(\mathbf{z}), & k = 1, \dots, k^* \\ \alpha_k + s_{\tau_k}(\boldsymbol{\xi}), & k = k^* + 1, \dots, K - 1 \end{cases}$$

and

$$\beta_k := \begin{cases} \sum_{j=k+1}^K (\alpha_j + s_{\tau_j}(\boldsymbol{\xi}) + s_{\tau_j}(\mathbf{z})), & k = 1, \dots, k^* \\ \sum_{j=k+1}^K (\alpha_j + s_{\tau_k}(\boldsymbol{\xi})), & k = k^* + 1, \dots, K - 1. \end{cases}$$

PROOF. See Appendix D in the supplementary material.

Here, we underline that we essentially derived an alternative construction of the Generalized Dirichlet distribution. Assuming that two vectors of weights share some common elements, and independent Dirichlet prior distributions are assigned to the free parameters of these weights, the posterior distribution of the first free parameter vector is a Generalized Dirichlet. Finally, notice that if  $\mathbf{v} = \emptyset$  (this is the case when the corresponding elements of the weights of the two mixtures are all equal to each other), the Generalized distribution (15) reduces to the distribution  $\mathcal{D}_{K-1}(\{\alpha_k + s_k(\boldsymbol{\xi}) + s_k(\mathbf{z}); k = 1, \dots, K\})$ , as expected, since in such a case  $(\mathbf{x}, \mathbf{y})$  forms a random sample of size  $r + s$  from the same population. On the other hand, if all weights are different, the full conditional distribution of  $\mathbf{u}, \mathbf{v}$  becomes a product of two independent Dirichlet distributions, as expected. Next we show that we can integrate out the parameters related to transcript expression and directly sample from the marginal posterior distribution of  $\boldsymbol{\xi}, \mathbf{z}, c | \mathbf{x}, \mathbf{y}$ .

**THEOREM 2.** *Integrating out the transcript expression parameters  $\mathbf{u}, \mathbf{v}$ , the full conditional distributions of allocation variables are written as:*

$$f(\boldsymbol{\xi}, \mathbf{z} | \mathbf{x}, \mathbf{y}, c) \propto \frac{\Gamma\left(\sum_{k \in C_1} \tilde{\alpha}_k + s_k(\boldsymbol{\xi}) + s_k(\mathbf{z})\right)}{\Gamma\left(\sum_{k \in C_1} \tilde{\alpha}_k + s_k(\boldsymbol{\xi})\right) \Gamma\left(\sum_{k \in C_1} \gamma_{\ell(k)} + s_k(\mathbf{z})\right)} \quad (17)$$

$$\times \prod_{k \in C_1} \Gamma(\tilde{\alpha}_k + s_k(\boldsymbol{\xi})) \Gamma(\gamma_{\ell(k)} + s_k(\mathbf{z}))$$

$$\times \prod_{k \in C_0} \Gamma(\tilde{\alpha}_k + s_k(\boldsymbol{\xi}) + s_k(\mathbf{z})) \prod_{i=1}^r f_{\xi_i}(x_i) \prod_{j=1}^s f_{z_j}(y_j)$$

$$P(\xi_i = k | \boldsymbol{\xi}_{[-i]}, \mathbf{z}, c, \mathbf{x}) \propto \begin{cases} (\tilde{\alpha}_k + s_k^{(i)}(\boldsymbol{\xi}) + s_k(\mathbf{z})) f_k(x_i), & k \in C_0 \\ \frac{\sum_{t \in C_1} \tilde{\alpha}_k + s_t^{(i)}(\boldsymbol{\xi}) + s_t(\mathbf{z})}{\sum_{t \in C_1} \tilde{\alpha}_t + s_t^{(i)}(\boldsymbol{\xi})} (\tilde{\alpha}_k + s_k^{(i)}(\boldsymbol{\xi})) f_k(x_i), & k \in C_1 \end{cases} \quad (18)$$

$$P(z_j = k | \mathbf{z}_{[-j]}, \boldsymbol{\xi}, c, \mathbf{y}) \propto \begin{cases} (\tilde{\alpha}_k + s_k(\boldsymbol{\xi}) + s_k^{(j)}(\mathbf{z})) f_k(y_j), & k \in C_0 \\ \frac{\sum_{t \in C_1} \tilde{\alpha}_t + s_t(\boldsymbol{\xi}) + s_t^{(j)}(\mathbf{z})}{\sum_{t \in C_1} \gamma_{\ell(t)} + s_t^{(j)}(\mathbf{z})} (\gamma_{\ell(k)} + s_k^{(j)}(\mathbf{z})) f_k(y_j), & k \in C_1 \end{cases} \quad (19)$$

where  $\tilde{\alpha}_k = \alpha_{\tau_k^{-1}}$ ,  $\ell(k) = \tau_k^{-1} - k^*$ ,  $s_k^{(i)}(\boldsymbol{\xi}) = \sum_{t \neq i} I(\xi_i = k)$ ,  $s_k^{(j)}(\mathbf{z}) = \sum_{t \neq j} I(z_i = k)$  for  $k = 1, \dots, K$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, s$ .

PROOF. See Appendix E in supplementary material.

Once again, note the intuitive interpretation of our model in the special cases where  $C_0 = \emptyset$  or  $C_1 = \emptyset$ . If  $C_0 = \emptyset$  (all transcripts are DE) then the nominator at the first line of Equation (17) becomes equal to  $\Gamma(\sum_k \alpha_k + r + s)$ , that is, independent of  $\boldsymbol{\xi}, \mathbf{z}$ . Hence, (17) reduces to the conditional distribution of the allocation variables when independent Dirichlet prior distributions are imposed to the mixture weights. On the contrary, when  $C_1 = \emptyset$  (all transcripts are EE), the distribution reduces to the product appearing in the last row of Equation (17). This is the marginal distribution of the allocations when considering that  $(\mathbf{x}, \mathbf{y})$  arise from the same population and after imposing a Dirichlet prior on the weights, as expected.

#### 2.4. MCMC samplers

In this section we consider the problem of sampling from the posterior distribution of the model in (12). We propose two (alternative) MCMC sampling schemes, depending on whether the transdimensional random variable  $\mathbf{v}$  is updated before or after  $c$ .

Note that given  $c$  everything has fixed dimension. However, as  $c$  varies on the set of its possible values, then  $\mathbf{v} \in \cup_{k \in \{0, 2, \dots, K\}} \mathcal{P}_{k-1}$ . This means that whenever  $c$  is updated,  $\mathbf{v}$  should change dimension. In order to construct a sampler that switches between different dimensions, a Reversible Jump MCMC method (Green, 1995) can be implemented (see also Richardson and Green (1997) and Papastamoulis and Iliopoulos (2009)). However, this step can be avoided since we have already shown that the transcript expression parameters can be integrated out. Thus, a collapsed sampler is also available. Given an initial state, the general work flow for the proposed samplers is the following (we avoid to explicitly state that all distributions appearing next are conditionally defined on the observed data  $\mathbf{x}, \mathbf{y}$ , although they should be understood as such).

##### rjMCMC Sampler

- (a) Update  $(\boldsymbol{\xi}, \mathbf{z}) | \boldsymbol{\theta}, \mathbf{w}$ .
- (b) Update  $(\mathbf{u}, \mathbf{v}) | c, \boldsymbol{\xi}, \mathbf{z}$ .
- (c) Update  $(\boldsymbol{\theta}, \mathbf{w}) | c, \tau, \mathbf{u}, \mathbf{v}$ .
- (d) Propose update of  $(c, \tau, \mathbf{v}) | \dots$
- (e) Update  $\pi | c$ .

##### Collapsed Sampler

- (a) Update  $\xi_i | \boldsymbol{\xi}_{[-i]}, \mathbf{z}, c, i = 1, \dots, r$ .
- (b) Update  $z_j | \boldsymbol{\xi}, \mathbf{z}_{[-j]}, c, j = 1, \dots, s$ .
- (c) Update a block of  $c | \boldsymbol{\xi}, \mathbf{z}$ .
- (d) Update  $\pi | c$ .
- (e) Update  $(\boldsymbol{\theta}, \mathbf{w}, \tau, \mathbf{u}, \mathbf{v}) | c, \boldsymbol{\xi}, \mathbf{z}$  (optional).

Note that step (e) is optional for the collapsed sampler. It is implemented only to derive the estimates of transcript expression but it is not necessary for the previous steps. The next paragraphs outline the workflow for step (d) of rjMCMC sampler and step (c) of the collapsed sampler. For full details the reader is referred to Appendices F and G in the Supplementary material.

*Reversible Jump sampler* Models of different dimensions are bridged using two move types, namely: “birth” and “death” of an index. The effect of a birth (death) move is to increase (decrease) the number of differentially expressed transcripts. These moves are complementary in the sense that the one is the reverse of the other. Note that this step proposes a candidate state which is accepted according to the acceptance probability.

*Collapsed sampler* In this case we randomly choose two transcripts ( $j_1$  and  $j_2$ ) and perform an update from the conditional distribution  $c_{j_1, j_2} | c_{-[j_1, j_2]} \xi, \mathbf{z}, \mathbf{x}, \mathbf{y}, \pi$ , which is detailed in Equations (G.1)–(G.4) in Section G of supplementary material. The random selection of the block  $\{j_1, j_2\} \subseteq \{1, \dots, K\}$  and the corresponding update of  $c_{j_1, j_2}$  from its full conditional distribution is a valid MCMC step because it corresponds to a Metropolis-Hastings step in which the acceptance probability equals 1 (see Lemma 2 in Appendix G of the supplementary material).

## 2.5. Clustering of reads and transcripts

In real RNA-seq datasets the number of transcripts could be very large. This imposes a great obstacle for the practical implementation of the proposed approach: the search space of the MCMC sampler consists of  $2^K$  elements (state vectors) and convergence of the sampler may be very slow. This problem can be alleviated by a cluster representation of aligned reads to the transcriptome. High quality mapped reads exhibit a sparse behaviour in terms of their mapping places: each read aligns to a small number of transcripts and there are groups of reads mapping to specific groups of transcripts. Hence, we can take advantage of this sparse representation of alignments and break the initial problem into simpler ones, by performing MCMC per cluster.

This clustering representation introduces an efficient way to perform parallel MCMC sampling by using multiple threads for transcript expression estimation. For this purpose we used the GNU parallel (Tange, 2011) tool, which effectively handles the problem of splitting a series of jobs (MCMC per cluster) into the available threads. The jobs are ordered according to the number of reads per cluster and the ones containing more reads

are queued first. GNU parallel efficiently spawns a new process when one finishes and keeps all available CPUs active, thus saving time compared to an arbitrary assignment of the same amount of jobs to the same number of available threads. For further details see Appendix H.

## 2.6. False Discovery Rate

Controlling the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995; Storey, 2003) is a crucial issue in multiple comparisons problems. Under a Bayesian perspective, any probabilistic model that defines a positive prior probability for DE and EE yields that  $\mathbb{E}(\text{FDR}|\text{data}) = \sum (1 - \hat{P}(c_k = 1|\mathbf{x}, \mathbf{y}))d_k/D$  (see for example Müller et al. (2004, 2006)), where  $d_k \in \{0, 1\}$  and  $D = \sum d_k$  denote the decision for transcript  $k$ ,  $k = 1, \dots, K$  and the total number of rejections, respectively. Consequently, FDR can be controlled at a desired level  $\alpha$  by choosing the transcripts that  $\hat{P}(c_k = 1|\mathbf{x}, \mathbf{y}) > 1 - \alpha$ , which is also the approach proposed by Leng et al. (2013). We have found that this rule achieves small false discovery rates compared to the desired level  $\alpha$ , but sometimes results to small true positive rate.

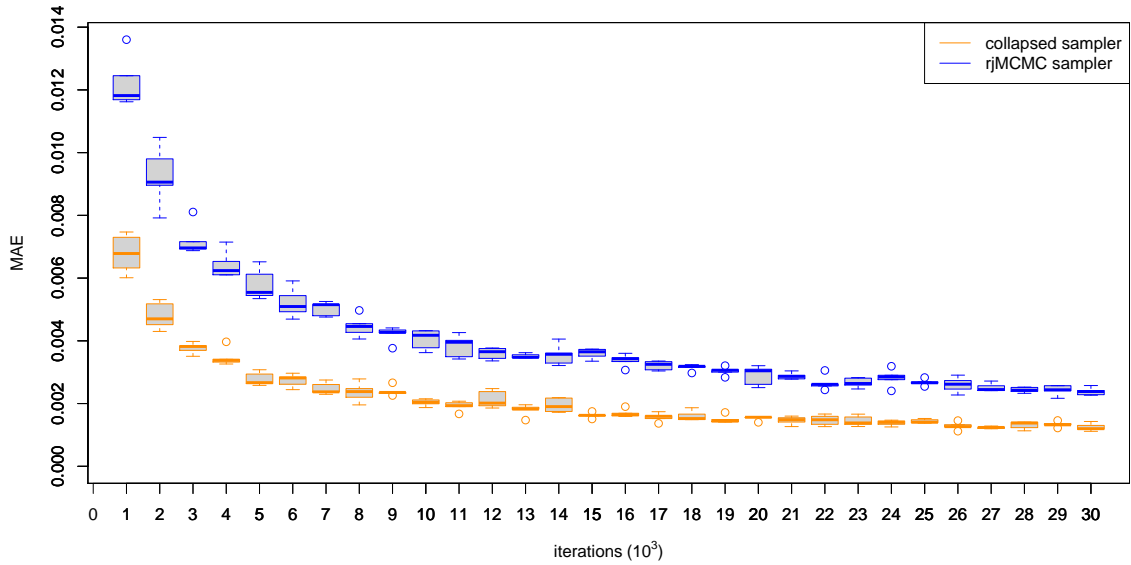
A less conservative choice is the following. Let  $q_1 \geq \dots \geq q_K$  denote the ordered values of  $\hat{P}(c_k = 1|\mathbf{x}, \mathbf{y})$ ,  $k = 1, \dots, K$  and define  $G_k := \frac{\sum_{j=1}^k (1 - q_j)}{k}$ ,  $k = 1, \dots, K$ . For any given  $0 < \alpha < 1$ , consider the decision rule:

$$d_k = \begin{cases} 1, & 1 \leq k \leq g \\ 0, & g + 1 \leq k \leq K \end{cases} \quad (20)$$

where  $g := \max\{k = 1 \dots, K : G_k \leq \alpha\}$ . It is quite straightforward to see that (20) controls the Expected False Discovery Rate at the desired level  $\alpha$ , since by direct substitution we have that

$$\mathbb{E}(\text{FDR}|\text{data}) = \frac{\sum_{k=1}^K (1 - \hat{P}(c_k = 1|\mathbf{x}, \mathbf{y}))d_k}{D} = \frac{\sum_{k=1}^g (1 - q_k)}{g} \leq \alpha.$$

An alternative is to use a rule optimizing the posterior expected loss of a predefined loss function. For example, the threshold  $c/(c+1)$  is the optimal cutoff under the loss function  $L = c\bar{\text{F}}\text{D} + \bar{\text{F}}\text{N}$ , where  $\bar{\text{F}}\text{D}$  and  $\bar{\text{F}}\text{N}$  denote the posterior expected counts of false discoveries and false negatives, respectively. Note that  $L$  is an extension of the  $(0, 1, c)$  loss functions for traditional hypothesis testing (Lindley, 1971), while a variety of alternative loss functions can be devised as discussed in Müller et al. (2004).



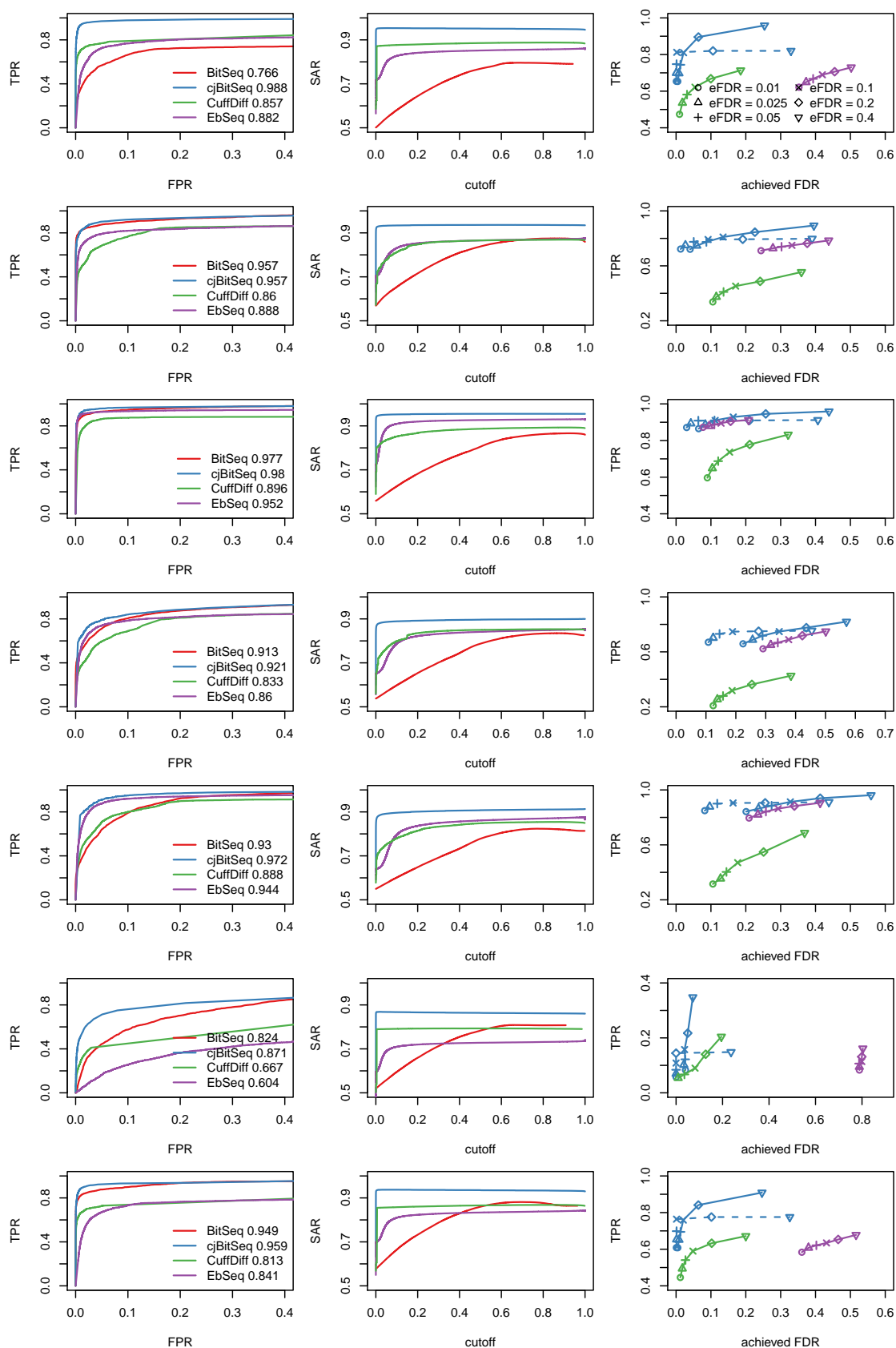
**Fig. 3.** Convergence of the ergodic means of posterior probabilities of DE for a toy example of  $K = 630$  transcripts. The “ground truth” for the posterior mean estimates ( $\hat{P}_g(c_k = 1)$ ;  $k = 1, \dots, K$ ) of these probabilities was inferred by running each sampler for 500000 iterations. Then, each sampler ran for a smaller number of  $m$  iterations resulting to the posterior mean estimates  $\hat{P}_m(c_k = 1)$ ;  $k = 1, \dots, K$ , for  $m = 1000, 2000, \dots, 30000$ . Finally, the averaged Mean Absolute Error of the posterior mean estimates was computed as:  $\frac{1}{K} \sum_{k=1}^K |\hat{P}_m(c_k = 1) - \hat{P}_g(c_k = 1)|$ . The boxplots correspond to five replications of the previous procedure.

### 3. Results

A set of simulation studies is used to benchmark the proposed methodology using synthetic RNA-seq reads from the *Drosophila melanogaster* transcriptome. The Spanki software (Sturgill et al., 2013) is used for this purpose. In addition to the simulated data study we also perform a comparison for two real datasets: a low and high coverage sequencing experiment using human data and a dataset from drosophila. In all cases, the reads are mapped to the reference transcriptome using Bowtie (version 2.0.6), allowing up to 100 alignments per read. Tophat (version 2.0.9) is also used for Cufflinks.

#### 3.1. Evaluation of samplers

We used a simulated dataset from  $K = 630$  transcripts (more details are described in Appendix H) and compare the posterior mean estimates between short and long runs. As shown in Figure 3, the collapsed sampler exhibits faster convergence than the rjMCMC sampler, hence in what follows we will only present results corresponding to the collapsed sampler. The reader is referred to the supplementary material (Appendices J and K) for



**Fig. 4.** Receiver Operating Characteristic (a), SAR measure (b) and Power-to-achieved-FDR (c) curves for scenario 1-7 (1st-7th row). The blue dashed lines correspond to the filtered cjBitSeq output by discarding transcripts with absolute log2 fold change less than 1.



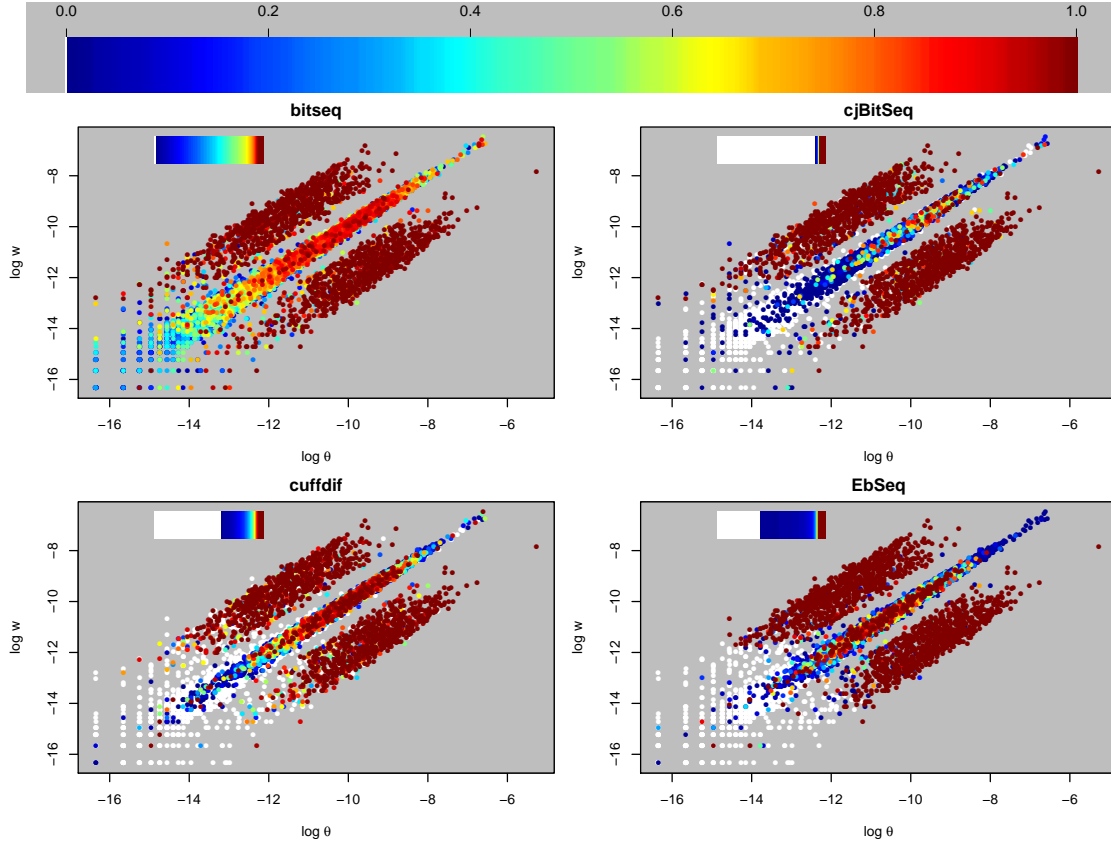
further comparisons (including autocorrelation function estimation and prior sensitivity) between our two MCMC schemes.

### 3.2. *Simulated data*

The input of the Spanki simulator is a set of reads per kilobase (rpkm) values per sample. This file is provided under a variety of different generative scenarios. Given the input files, Spanki simulates RNA-seq reads (in fastq format) according to the specified rpkm values. Seven scenarios are used to generate the data: two Poisson replicates per condition (scenario 1), three Negative Binomial replicates per condition (scenario 2), 9 Negative Binomial replicates (scenario 3), three Negative Binomial replicates per condition with five times higher variability among replicates compared to scenario 2 (scenario 4), same variability with scenario 4 but a smaller range for the mean rpkm values (scenario 5). The last two scenarios are revisions of the first scenario with smaller fold changes (scenario 6) and large differences in the number of reads between conditions (scenario 7). See supplementary Figure 9 and Appendix K for the details of the ground truth used in our simulations.

Next, we applied the proposed method and compared our results against Bitseq, Cuffdiff and EBSeq, using (a) ROC, (b) SAR-measure (Sing et al., 2005) and (c) Power-to-achieved-FDR curves, as shown in Figure 4. For the comparison in (c) the FDR decision of our model is based on the rule (20). Moreover, only methods that control the FDR are taken into account in (c), hence BitSeq Stage 2 is excluded. In addition to this FDR control procedure, we also provide adjusted rates after imposing a threshold to the log-fold change of the cjBitSeq sampler: all transcripts with estimated absolute log2 fold change less than 1 are filtered out (results correspond to the blue dashed line). A typical behaviour of the compared methods is illustrated in Figure 5, displaying true expression values used in Scenario 3. We conclude that our method infers an almost ideal classification, something that is not the case for the other methods despite the large number of replicates used.

In order to summarize our findings, Figure 6 displays the complementary area under the curve for each scenario. Averaging across all simulation scenarios, we conclude that our method is almost 2 times better than BitSeq Stage 2, 3 times better than EBSeq and 3.2 times better than Cuffdiff. Finally, we compare the estimated relative abundance of transcripts against the true values used to generate the data, using the average across all replicates of a given condition. Figure 6 (bottom) displays the Mean Absolute Error between the logarithm of true transcript expression and the corresponding estimates ac-



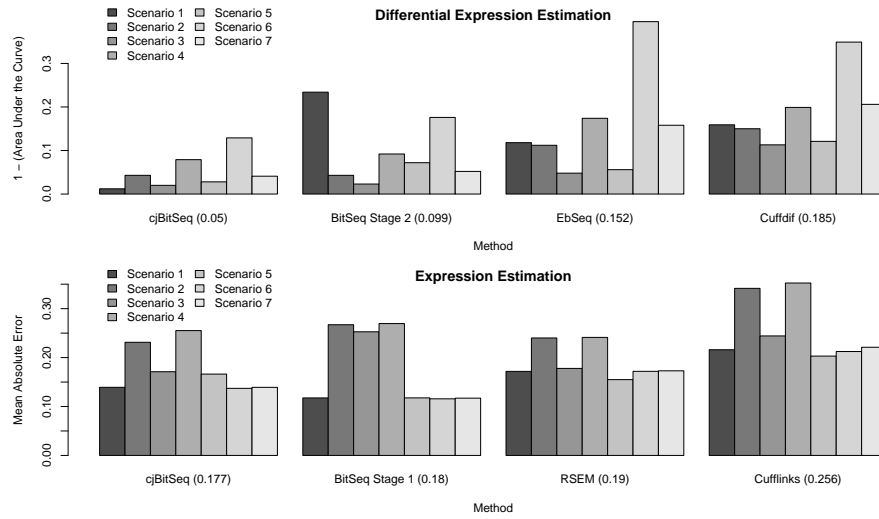
**Fig. 5.** True log-relative expression values for Scenario 3 (average of nine replicates per condition,  $\approx 24$  million reads in total). The color corresponds to the evidence of differential expression according to each method and the legend shows the relative frequency of colors.

cording to each method. We see that cjBitSeq, BitSeq stage 1 and RSEM exhibit a similar behaviour, while all performing significantly better than Cufflinks. Although there is no consistent ordering among the first three methods, averaging across all experiments we conclude that cjBitSeq is ranked first.

We have also tested the sensitivity of our method with respect to the prior distributions of differential expression (3) by setting  $\pi = 0.5$  (see supplementary Figure 11 and the corresponding discussion in Appendix K). We conclude that the prior distribution does not affect the ranking of methods both for differential and expression estimation.

### 3.3. Human data

This example demonstrates the proposed algorithm to differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOXA1, see Trapnell et al. (2013) for full details. There are three biological replicates in the two conditions.



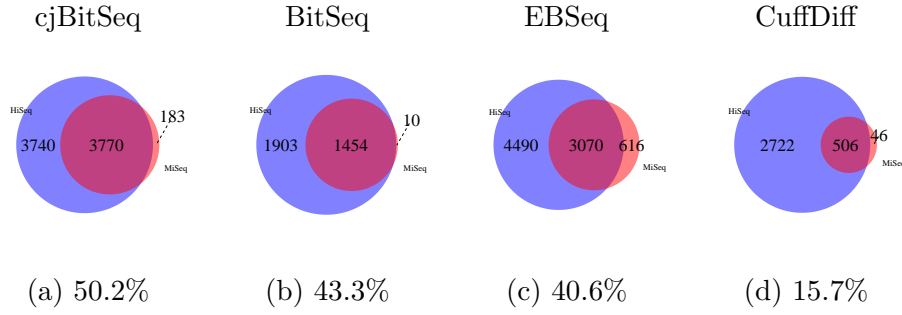
**Fig. 6.** Simulated data: Ranking of methods with respect to estimation of differential expression (top) and the log of relative expression (bottom). The methods are ordered according to the averaged complementary Area Under the Curve and Mean Absolute Error (shown in parenthesis).

The experiment is carried out using two sequencing platforms: HiSeq and MiSeq, where MiSeq produced only 23% of the number of reads in the HiSeq data. Here, these reads are mapped to hg19 (UCSC annotation) using Bowtie 2, consisting of  $K = 48009$  transcripts. In total, there are 96969106 and 21271542 mapped reads for HiSeq and MiSeq sequencers, respectively. Trapnell et al. (2013) demonstrated the ability of Cuffdiff2 to recover the transcript dynamics from the HOXA1 knockdown when using the significantly smaller amount of data generated by MiSeq compared to HiSeq.

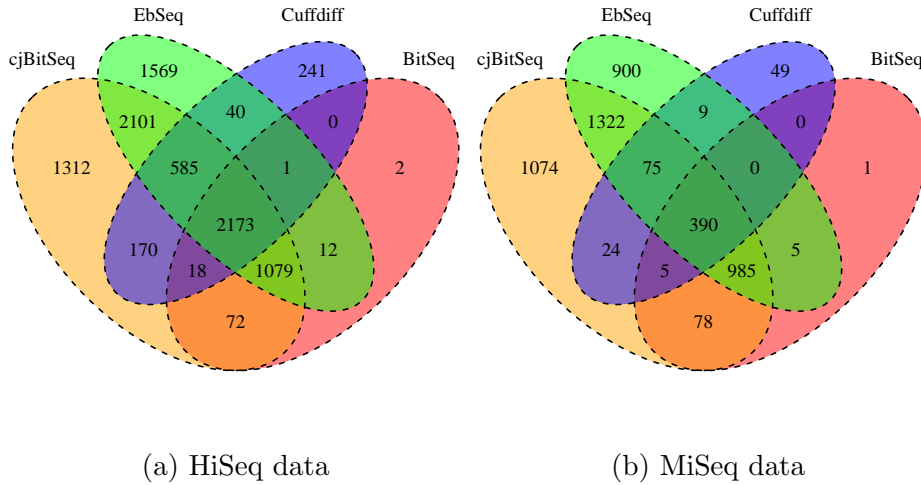
Applying cjBitSeq to the MiSeq data recovers 50.2% of the DE transcripts from HiSeq. On the other hand, there are 183 transcripts reported as DE with the MiSeq data but not the HiSeq data (Figures 8(a) and 8(b)). The corresponding percentages for BitSeq stage 2, EBSeq and Cuffdiff are 43.3%, 40.6% and 15.7%, respectively (see Figures 8(b), 8(c) and 8(d)). We conclude that the proposed model returns the largest proportion of consistently DE transcripts between platforms. The number of transcripts which are simultaneously reported as DE is equal to 2173 and 390 for HiSeq and MiSeq data, respectively (see Figures 7.a and 7.b). Finally, cjBitSeq and EBSeq provide the most highly correlated classifications (see Table 1 of supplementary material).

#### 4. Discussion

We have proposed a probabilistic model for the simultaneous estimation of transcript expression and differential expression between conditions. Building upon the BitSeq frame-



**Fig. 7.** HOXA1 knockdown dataset: Significant transcript list returned by cjBitSeq (a), BitSeq (b), EBSeq (c) and CuffDiff (d) when using HiSeq (blue) and MiSeq (red) data. FDR for cjBitSeq, EBSeq and CuffDiff set to 0.05, while for BitSeq:  $PPLR < 0.025$  or  $PPLR > 0.975$



**Fig. 8.** HOXA1 knockdown dataset: Contiguity of methods when using HiSeq (a) and MiSeq (b) data. FDR for cjBitSeq, EBSeq and CuffDiff set to 0.05, while for BitSeq:  $PPLR < 0.025$  or  $PPLR > 0.975$

work, the new Bayesian hierarchical model is conjugate for fixed dimension variables. A by-product is a new interpretation of the Generalized Dirichlet distribution, which naturally appears in (15) as the full conditional distribution of a random variable describing one of the free parameters corresponding to two proportion vectors under the constraint that some of the weights are equal to each other. We implemented two MCMC samplers, a reversible-jump and collapsed Gibbs sampler, and we found the collapsed Gibbs sampler to converge faster. To greatly reduce the dimensionality of the parameter space for inference we developed a transcript clustering approach which allows inference to be carried out independently on subsets of transcripts that share aligned reads. According to Lemma 3 in the supplementary material (Appendix H), this clustered version of the

vanilla algorithm converges to the proper marginal distribution for each cluster. Thus, the algorithm has the nice property that it can be run in parallel for each cluster, while the memory requirements are quite low, providing a simple parallelisation option.

The applications to simulated and real RNA-seq data reveals that the proposed method is highly competitive with the current state of the art software dealing with DE analysis at the transcript level. Note that the simulated data was generated under a variety of different scenarios and including different levels of replication and biological variation. We simulated transcript RPK values with variability following either the Poisson or the Negative Binomial distribution with various levels for the dispersion around the mean. We conclude that our method is quite robust in expression estimation and in classifying transcripts as DE or not. Compared to standard two-stage pipelines it is ranked as the best method under a wide range of generative scenarios.

RNA-seq data are usually replicated such that there is more than one datasets available for each condition. In such a way, biological variability between repetitions of the same experiment can be taken into account. The amount of variability between replicates can be quite high depending on the experimental conditions. Two-stage approaches for estimating differential expression are strongly focused on modelling this inter-replicate variability. This is not the case for our method at present and all replicates of a given condition are effectively pooled together prior to inference. Modelling the variability between replicates would significantly increase the complexity of our approach as it is technically challenging to retain conjugacy. However, according to our simulation studies, we have found that pooling replicates together and jointly estimating expression and differential expression balances the loss through ignoring variability between replicates in many cases. Nevertheless, an extension to also model inter-replicate variability would be very interesting and could be expected to improve performance when there is high inter-replicate dispersion.

The proposed method was developed focusing to the comparison of two conditions and its extension to more general settings is another interesting area for future research. A remarkable property of the parameterization introduced in Equations (5) and (6) is that its extension is straightforward when  $J > 2$ : it can be shown that in this case there is one parameter of constant dimension and  $J - 1$  parameters of varying dimension. Let  $\mathbf{u} = \mathbf{u}^{(1)}$  be the vector of relative abundances for condition 1. For a given condition  $j = 2, \dots, J$  define a vector  $\mathbf{v}_j$  containing the expression of transcripts not being equal to any of the previous conditions  $1, \dots, j - 1$ . Note that  $\mathbf{v}_j$  is a random variable with varying length (between 0 and  $K$ ). Furthermore, for  $j \geq 2$  define the vectors  $\mathbf{u}_k^{(j)}$ ,  $k = 1, \dots, j - 1$ ,

containing the expression of transcripts shared with condition  $k$  but not with  $1, \dots, k-1$ . It follows that  $\mathbf{u}_k^{(j)}$  can be written as a function of  $\mathbf{u}^{(1)}$  and  $\mathbf{v}_k$ ,  $k = 1, \dots, j-1$ . Hence, the relative transcript expression vector for condition  $j$  can be expressed as a suitable permutation of  $(\mathbf{u}_1^{(j)}, \dots, \mathbf{u}_{j-1}^{(j)}, \mathbf{v}_j)$ . However, the question of whether the proposed model stays conjugate for fixed dimension updates remains an open problem. If yes, the design of more sophisticated move-types between different models would be also crucial to the convergence of the algorithm since the search space is increased.

The source code of the proposed algorithm is compiled for LINUX distributions and it is available at <https://github.com/mqbssppe/cjBitSeq>. The simulation pipeline is available at [https://github.com/ManchesterBioinference/cjBitSeq\\_benchmarking](https://github.com/ManchesterBioinference/cjBitSeq_benchmarking). Cluster discovery and MCMC sampling is coded in R and C++, respectively. Parallel runs of the MCMC scheme are implemented using the GNU parallel (Tange, 2011) shell tool. The computing times needed for our datasets are reported in supplementary Table 2.

## Acknowledgments

The research was supported by MRC award MR/M02010X/1, BBSRC award BB/J009415/1 and EU FP7 project RADIANT (grant 305626). The authors acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester. We also thank the editor and two anonymous reviewers for their helpful comments and suggestions which helped us to improve the manuscript.

**Supplementary material:** We provide the proofs of our Lemmas and Theorems, a detailed description of the reversible jump proposal and the Gibbs updates of state vector of the collapsed sampler. Also included are details of alignment probabilities and some useful properties of the Generalized Dirichlet distribution. We also perform various comparisons between the rjMCMC and collapsed samplers and examine their prior sensitivity. Finally we describe the generative schemes for the simulation study and some guidelines for the practical implementation of the algorithm.

## References

- Anders, A. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

- Connor, R. J. and Mosimann, J. E. (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, **64**, 194–206.
- Gelfand, A. and Smith, A. (1990) Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, **85**, 398–409.
- Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721–741.
- Glaus, P., Honkela, A. and Rattray, M. (2012) Identifying differentially expressed transcripts from RNA-Seq data with biological variation. *Bioinformatics*, **28**, 1721–1728.
- Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Gu, J., Wang, X., Halakivi-Clarke, L., Clarke, R. and Xuan, J. (2014) BADGE: A novel Bayesian model for accurate abundance quantification and differential analysis of RNA-Seq data. *BMC Bioinformatics* 2014, **15**.
- Hensman, J., Papastamoulis, P., Glaus, P., Honkela, A. and Rattray, M. (2015) Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, **31**, 3881–3889.
- Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **186**, 453–461.
- Jia, C., Guan, W., Yang, A., Xiao, R., Tang, W. H. W., Moravec, C. S., Margulies, K. B., Cappola, T. P., Li, C. and Li, M. (2015) MetaDiff: differential isoform expression analysis using random-effects meta-regression. *BMC Bioinformatics*, **16**, 1–12.
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**.
- Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M. and Kendzierski, C. (2013) EBSeq: An empirical Bayes hierarchical model for inference in RNA-Seq experiments. *Bioinformatics*.

- Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
- Lindley, D. V. (1971) *Making Decisions*. Willey, New York.
- Mortazavi, A., Williams, B., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621–628.
- Müller, P., Parmigiani, G. and Rice, K. (2006) FDR and Bayesian multiple comparisons rules. *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics*.
- Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004) Optimal sample size for multiple testing. *Journal of the American Statistical Association*, **99**, 990–1001.
- Nariai, N., Hirose, O., Kojima, K. and Nagasaki, M. (2013) TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. *Bioinformatics*, **18**, 2292–2299.
- Nicolae, M., Mangul, S., Mandoiu, I. and Zelikovsky, A. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, **6**:9.
- Papastamoulis, P., Hensman, J., Glaus, P. and Rattray, M. (2014) Improved variational Bayes inference for transcript expression estimation. *Statistical applications in Genetics and Molecular Biology*, **13**, 213–216.
- Papastamoulis, P. and Iliopoulos, G. (2009) Reversible jump MCMC in mixtures of normal distributions with the same component means. *Computational Statistics and Data Analysis*, **53**, 900–911.
- Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B*, **59**, 731–758.
- Robinson, M., McCarthy, D. and Smyth, G. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Rossell, D., Attolini, C, S.-O., Kroiss, M. and Stocker, A. (2014) Quantifying alternative splicing from paired-end RNA-Sequencing data. *Annals of Applied Statistics*, **8**, 309–330.
- Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCr: visualizing classifier performance in R. *Bioinformatics*, **21**, 7881.



- Storey, J. D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of statistics*, 2013–2035.
- Sturgill, J., Malone, J. H., Sun, X., Smith, H. E., Rabinow, L., Samson, M. L. and Oliver, B. (2013) Design of RNA splicing analysis null models for post hoc filtering of drosophila head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics*, **14**, 320.
- Tange, O. (2011) GNU parallel - the command-line power tool. *login: The USENIX Magazine*, **36**, 42–47. URL <http://www.gnu.org/s/parallel>.
- Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology*, **31**, 46–53.
- Trapnell, C., Pachter, L. and Salzberg, S. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.
- Wong, T. (1998) Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, **97**, 165–181.
- (2010) Parameter estimation for generalized Dirichlet distributions from the sample estimates of the first and the second moments of random variables. *Computational Statistics and Data Analysis*, **54**, 1756–1765.

## Appendices

### A. Alignment probability

In this section the component specific density (1) is defined. For single-end reads, let  $\ell_i > 0$  denotes the length of read  $x_i$ ,  $i = 1, \dots, n$ . Assume that  $x_i$  aligns at some position  $p$  of a given transcript  $k$ ,  $k = 1, \dots, K$  and that the corresponding transcript length equals to  $L_k > 0$ . Note that both  $L_k$ ,  $\ell_i$  are known quantities. The general form of observing this alignment equals to

$$f_k(x_i) = P(x_i = p|k) = \frac{b_k(p)}{\sum_{j=1}^{L_k-\ell_i+1} b_k(j)}, \quad (\text{A.1})$$

where  $b_k(j)$  denotes the bias for a particular position  $p$  on transcript  $k$ . In case of a Uniform read distribution, the previous equation reduces to:

$$f_k(x_i) = \frac{1}{L_k - \ell_i + 1}. \quad (\text{A.2})$$

More complex choices are also available. In particular, a separate variable length Markov is used to capture the position and sequence specific biases for the 5' and 3' ends of the fragment. For more details the reader is referred to Glaus et al. (2012).

In case of paired-end reads, the fragment length  $\ell$  is also taken into account. The fragment length distribution  $f(\ell|k)$  is assumed to be log-normal with parameters given by the user or estimated from read pairs with only a single transcript alignment. In this case the alignment probability of a read pair is given as

$$f_k(x_i) = f_k(x_i = p, \ell) = f(\ell|k) \frac{b_k(p)}{\sum_{j=1}^{L_k - \ell_i + 1} b_k(j)}. \quad (\text{A.3})$$

Finally, the alignment probabilities also take into account base-calling errors using the Phred score. For full details see Glaus et al. (2012). In our presented examples we assumed the Uniform read distribution.

The sampling scheme of the RNA-seq procedure for single-end reads is displayed in Figure 9. The four long sequences of letters correspond to transcripts which share specific parts of their sequence. The gray coloured regions are skipped, so each transcript is consisting only from the remaining region (coloured in red, blue, green and purple). The short reads are randomly generated sequences from each transcript. Note that most reads align to more than one transcript.

## B. The Generalized Dirichlet distribution

This generalization of the Dirichlet distribution was originally introduced by Connor and Mossiman (1969). The most prominent difference with a typical Dirichlet is that the Generalized Dirichlet family has a richer covariance structure. For example, only negative correlation between any pairs of variables is allowed under the Dirichlet distribution, while the Generalized Dirichlet can also allow positive correlation. Another difference is that any permutation of a vector of proportions which follows a Dirichlet distribution is also distributed as a Dirichlet distribution. However, this is not necessarily true for the Generalized Dirichlet distribution.

In this paper we follow the parameterization of the Generalized Dirichlet distribution introduced by Wong (1998). Let  $\mathbf{X} = (X_1, \dots, X_k; X_{k+1})$ , with  $\sum_{j=1}^k X_j \leq 1$ ,  $X_j \geq 0$



**Fig. 9.** Illustration of the RNA-seq sampling scheme using single reads and a small set of four transcripts (red, blue, green and purple). Gray color corresponds to skipped regions (exons). From each transcript we simulated 10 reads each one consisting of 10 base pairs, displayed under each transcript.

for  $j = 1, \dots, k$  and  $X_{k+1} = 1 - X_1 - \dots - X_k$ . Assume that  $\alpha_j > 0$ ,  $\beta_j > 0$  be a set of parameters,  $j = 1, \dots, k$ . Then,  $\mathbf{X} \sim \mathcal{GD}(\alpha_1, \dots, \alpha_k; \beta_1, \dots, \beta_k)$  if the probability density function is written as

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \prod_{j=1}^k \frac{x_j^{\alpha_j-1} (1-x_1-\dots-x_j)^{\gamma_j}}{B(\alpha_j, \beta_j)}, & \sum_{j=1}^k x_j \leq 1, x_j \geq 0, j = 1, \dots, k \\ 0, & \text{otherwise} \end{cases} \quad (\text{B.1})$$

where  $\gamma_j = \beta_j - \alpha_{j+1} - \beta_{j+1}$  for  $j = 1, \dots, k-1$ , and  $\gamma_k = \beta_k - 1$  and  $B(\cdot, \cdot)$  denotes the Beta function. Note that when

$$\beta_j = \alpha_{j+1} + \beta_{j+1}, \quad j = 1, \dots, k-1, \quad (\text{B.2})$$

a Generalized Distribution reduces to a standard Dirichlet distribution.

An important property of both Dirichlet and Generalized Dirichlet is that they can be constructed using a stick breaking process. The following result is from Connor and Mossiman (1969): Define  $\zeta_1 = X_1$  and  $\zeta_j = X_j/V_{j-1}$  for  $j = 2, 3, \dots, k$ , where  $V_j = 1 - X_1 - \dots - X_{j-1}$ . If  $\zeta_j \sim \text{Beta}(\alpha_j, \beta_j)$ , independent for  $j = 1, \dots, k$ . Hence we can

construct  $X$  as follows:

$$\begin{aligned} X_1 &= \zeta_1 \\ X_j &= \zeta_j(1 - X_1 - \dots - X_{j-1}) = \zeta_j \prod_{i=1}^{j-1} (1 - \zeta_i), j = 2, 3, \dots, k \\ X_{k+1} &= 1 - \prod_{i=1}^k (1 - \zeta_i). \end{aligned}$$

In this case:  $\mathbf{X} = (X_1, \dots, X_k; X_{k+1}) \sim \mathcal{GD}(\alpha_1, \dots, \alpha_k; \beta_1, \dots, \beta_k)$  (Connor and Mossiman, 1969). Notice that if  $\beta_j = \sum_{k=j+1}^{k+1} \alpha_k$  and also define  $\beta_{k+1} = \alpha_{k+1}$  for a given  $\alpha_{k+1} > 0$ , then  $\mathbf{X} \sim \mathcal{D}(\alpha_1, \dots, \alpha_k, \alpha_{k+1})$ .

The previously described construction is closely related to the notion of neutrality which was also introduced by Connor and Mossiman (1969): “a neutral vector of proportions do not influence the proportional division of the remaining interval among the remaining variables”. In particular: a vector of proportions is completely neutral if and only if  $\zeta_i$ ’s are mutually independent (Theorem 2, Connor and Mossiman, 1969). The concept of complete neutrality as well as the representation through the  $\zeta$  random variables characterize both the Dirichlet and Generalized Dirichlet distributions and it will be useful for the proof of Theorem 1.

### C. Proof of Theorem 1

We start with the derivation of the marginal distribution of  $\boldsymbol{\theta}$ . According to (5), for any given state vector  $c$ ,  $\boldsymbol{\theta}$  can be expressed as a suitable permutation of a random variable  $\mathbf{u} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$ . Thus, we can write that:

$$f(\boldsymbol{\theta}) = \sum_{c \in \mathcal{C}} P(c) f(\tau_c^{-1} \mathbf{u}). \quad (\text{C.1})$$

Now recall that any permutation of  $\mathbf{u}$  is also distributed according to a Dirichlet distribution and its parameters are just the corresponding permutation of the initial parameters. This means that  $\tau_c^{-1} \mathbf{u} \sim \mathcal{D}_{K-1}(\tau_c^{-1} \boldsymbol{\alpha})$ , where  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ . Hence, in the general case where  $\boldsymbol{\alpha}$  is an arbitrary vector of strictly positive numbers, (C.1) is a mixture of Dirichlet distributions. Now notice that if  $\alpha_k = \alpha > 0$ , for all  $k = 1, \dots, K$ , then  $\tau_c^{-1} \boldsymbol{\alpha} = \boldsymbol{\alpha}$  for  $c \in \mathcal{C}$  and (C.1) reduces to  $\mathcal{D}_{K-1}(\boldsymbol{\alpha})$ .

The analogous result for  $\mathbf{w}$  demands a little bit more effort. At first notice that for any given  $c$ ,  $\mathbf{w}$  can be expressed according to Equation (6) as a suitable permutation of

$$\boldsymbol{\rho} = (u_1, \dots, u_{k^*}, v_1 D_c, \dots, v_{c^+} D_c),$$

where  $D_c = \sum_{k=k^*+1}^K u_k$ . Following the similar argument with  $\boldsymbol{\theta}$ , it will be sufficient to prove that  $\boldsymbol{\rho}$  follows a Dirichlet distribution. From the discussion in Appendix B, it is equivalent to establish that  $\boldsymbol{\rho}$  is completely neutral with  $\zeta_k \sim \text{Beta}(\delta_j, \sum_{k=j+1}^K \delta_k)$  independent for  $k = 1, \dots, K-1$  for some  $\delta_k > 0$ ,  $k = 1, \dots, K$ .

Let us define now the following variables:

$$\begin{aligned}\zeta_1 &= u_1 \\ \zeta_2 &= \frac{u_2}{1 - u_1} \\ &\vdots \\ \zeta_{k^*} &= \frac{u_{k^*}}{1 - u_1 - \dots - u_{k^*-1}} \\ \zeta_{k^*+1} &= v_1 \\ \zeta_{k^*+2} &= \frac{v_2}{1 - v_1} \\ &\vdots \\ \zeta_{K-1} &= \frac{v_{K-1}}{1 - v_1 - \dots - v_{K-2}}.\end{aligned}$$

Since  $\mathbf{u}$  and  $\mathbf{v}$  are independent and distributed according to (9) and (10) it follows that  $\zeta_k \sim \text{Beta}(\alpha_k, \sum_{j=k+1}^K \alpha_j)$  for  $k = 1, \dots, k^*$  and  $\zeta_{k^*+\ell} \sim \text{Beta}(\gamma_\ell, \sum_{j=\ell+1}^{c+} \gamma_j)$  for  $\ell = 1, \dots, c_+$ . Furthermore,  $\zeta_k$  are mutually independent for  $k = 1, \dots, K-1$ .

Now, observe that  $\rho_1 = \zeta_1$ ,  $\rho_k = \frac{\zeta_k}{1 - \rho_1 - \dots - \rho_{k-1}}$ ,  $k = 2, \dots, K-1$  and  $\rho_K = 1 - \sum_{j=1}^{K-1} \rho_j$ . But  $\zeta$ 's are mutually independent and Beta distributed, consequently  $\boldsymbol{\rho}$  follows a Generalized Dirichlet distribution:

$$\boldsymbol{\rho} \sim \mathcal{GD} \left( \alpha_1, \dots, \alpha_{k^*}, \gamma_1, \dots, \gamma_{c_+}; \sum_{j=2}^K \alpha_j, \dots, \sum_{j=k^*+1}^K \alpha_j, \sum_{j=2}^{c_+} \gamma_j, \dots, \gamma_{c_+} \right). \quad (\text{C.2})$$

Since  $\mathbf{w} = \tau_c^{-1} \boldsymbol{\rho}$  for any given  $c$ , in general, the marginal prior distribution of  $\mathbf{w}$  is a mixture of permutations of Generalized Dirichlet distributions (as previously discussed, the Generalized Dirichlet distribution is not permutation invariant). In the special case that  $\alpha_k = \gamma_k = \alpha > 0$  for all  $k = 1, \dots, K$ , the property (B.2) implies that the distribution (C.2) reduces to  $\mathcal{D}(\boldsymbol{\alpha})$ . The result follows using the same argument as the one used for  $\boldsymbol{\theta}$ .

## D. Proof of Lemma 2

From (12) we have that:

$$\mathbf{u}, \mathbf{v} | \dots \propto \prod_{i=1}^r \theta(\tau, \mathbf{u})_{\xi_i} \prod_{j=1}^s \mathbf{w}(\tau, \mathbf{u}, \mathbf{v})_{z_j} \prod_{k=1}^K u_k^{\alpha_k-1} \prod_{\ell=1}^{c_+} v_\ell^{\gamma_\ell-1}$$

$$\begin{aligned}
& \propto \prod_{k=1}^K \theta(\tau, \mathbf{u})_k^{s_k(\boldsymbol{\xi})} \prod_{k=1}^K \mathbf{w}(\tau, \mathbf{u}, \mathbf{v})_k^{s_k(\mathbf{z})} \prod_{k=1}^K u_k^{\alpha_k-1} \prod_{\ell=1}^{c_+} v_\ell^{\gamma_\ell-1} \\
& \propto \prod_{k=1}^K \tau^{-1} u_k^{s_k(\boldsymbol{\xi})} \prod_{k \in C_0(c)} \mathbf{w}(\tau, \mathbf{u})_k^{s_k(\mathbf{z})} \prod_{k \in C_1(c)} \mathbf{w}(\tau, \mathbf{u}, \mathbf{v})_k^{s_k(\mathbf{z})} \\
& \quad \times \prod_{k=1}^K u_k^{\alpha_k-1} \prod_{\ell=1}^{c_+} v_\ell^{\gamma_\ell-1} \\
& \propto \prod_{k=1}^K u_k^{s_{\tau_k}(\boldsymbol{\xi})} \prod_{k=1}^{k^*} u_k^{s_{\tau_k}(\mathbf{z})} \prod_{k=k^*+1}^K \left( v_{k-k^*} \sum_{j=k^*+1}^K u_j \right)^{s_{\tau_k}(\mathbf{z})} \\
& \quad \times \prod_{k=1}^K u_k^{\alpha_k-1} \prod_{\ell=1}^{c_+} v_\ell^{\gamma_\ell-1} \\
& \propto \prod_{k=1}^{k^*} u_k^{\alpha_k + s_{\tau_k}(\boldsymbol{\xi}) + s_{\tau_k}(\mathbf{z}) - 1} \prod_{k=k^*+1}^K u_k^{\alpha_k + s_{\tau_k}(\boldsymbol{\xi}) - 1} \left( \sum_{j=k^*+1}^K u_j \right)^{\sum_{j=k^*+1}^K s_{\tau_j}(\mathbf{z})} \\
& \quad \times \prod_{\ell=1}^{c_+} v_\ell^{\gamma_\ell + s_{\tau_{\ell+k^*}}(\mathbf{z}) - 1} \tag{D.1}
\end{aligned}$$

The last expression yields to conditional independence of  $\mathbf{u}$  and  $\mathbf{v}$ . Moreover, it is straightforward to see that the full conditional distribution of  $\mathbf{v}$  is the one defined in expression (16). The easiest way to see that the conditional distribution of  $\mathbf{u}$  is the one defined in (15) is to evaluate the density function (B.1) with the parameters given in (15), make all simplifications and then end up to the first row of last equation.

Finally, it is important to stress here the convenience of defining  $\mathbf{u}$  in a way that the set of Equally Expressed transcripts ( $C_0$ ) is followed by the set of Differentially Expressed transcripts ( $C_1$ ), as well as the permutation of the indices as in Definition 2. Note that the term corresponding to  $\sum_{j=k^*+1}^K u_j$  in expression D.1 refers to the sum of weights of the Differentially Expressed transcripts. If  $C_1$  would be a random subset of indices and not the one corresponding to the last  $c_+ = K - k^*$  ones, then it would not be possible to directly express the first line of D.1 as a member of the Generalized Dirichlet family, but rather as a permutation of a Generalized Dirichlet distributed random variable.

## E. Proof of Theorem 2

Let  $\mathcal{A}_c = \mathcal{P}_{K-1} \times \mathcal{P}_{c_+-1}$  and also note that when  $c_+ = 0$  then  $\mathcal{A}_c$  reduces to  $\mathcal{P}_{K-1}$ . From Equation (12) and Lemma 2 we have that:

$$f(\boldsymbol{\xi}, \mathbf{z} | \mathbf{x}, \mathbf{y}, c) \propto \int_{\mathcal{A}_c} H(\mathbf{u}, \mathbf{v}, \boldsymbol{\xi}, \mathbf{z}, c) d\mathbf{u} d\mathbf{v} \prod_{i=1}^r f_{\xi_i}(x_i) \prod_{j=1}^s f_{z_j}(y_j), \tag{E.1}$$

where  $H(\mathbf{u}, \mathbf{v}, \boldsymbol{\xi}, \mathbf{z}, c)$  denotes the expression (D.1). Now recall that according to Lemma 2, the full conditional distribution of  $\mathbf{u}, \mathbf{v} | \dots$  becomes a product of independent Generalized Dirichlet and Dirichlet distributions. This means that

$$\begin{aligned} \int_{\mathcal{A}_c} H(\mathbf{u}, \mathbf{v}, \boldsymbol{\xi}, \mathbf{z}, c) d\mathbf{u} d\mathbf{v} &= \prod_{k=1}^{K-1} B(\lambda_k, \beta_k) \frac{\prod_{\ell=1}^{c_+} \Gamma(\gamma_\ell + s_{\tau_{\ell+k^*}}(\mathbf{z}))}{\Gamma(\sum_{\ell=1}^{c_+} \gamma_\ell + s_{\tau_{\ell+k^*}}(\mathbf{z}))} \\ &= \prod_{k=1}^{K-1} B(\lambda_k, \beta_k) \frac{\prod_{k \in C_1} \Gamma(\gamma_{\tau_k^{-1}-k^*} + s_k(\mathbf{z}))}{\Gamma(\sum_{k \in C_1} \gamma_{\tau_k^{-1}-k^*} + s_k(\mathbf{z}))}. \end{aligned} \quad (\text{E.2})$$

Define  $\beta_0 = \sum_{j=1}^K \alpha_j + r + s$ . Observe that  $\beta_k + \lambda_k = \beta_{k-1}$  for all  $k \neq k^* + 1$ . Now simplify the product of Beta functions as follows:

$$\begin{aligned} \prod_{k=1}^{K-1} B(\lambda_k, \beta_k) &= \prod_{k=1}^{K-1} \frac{\Gamma(\lambda_k) \Gamma(\beta_k)}{\Gamma(\lambda_k + \beta_k)} \\ &= \frac{\left( \prod_{k=1}^{k^*} \Gamma(\lambda_k) \right) \Gamma(\beta_{k^*})}{\Gamma(\beta_0)} \frac{\left( \prod_{k=k^*+1}^{K-1} \Gamma(\lambda_k) \right) \Gamma(\beta_{K-1})}{\Gamma(\beta_{k^*+1} + \lambda_{k^*+1})} \\ &= \frac{\Gamma(\beta_{k^*}) \Gamma(\beta_{K-1})}{\Gamma(\beta_0) \Gamma(\beta_{k^*+1} + \lambda_{k^*+1})} \prod_{k=1}^{K-1} \Gamma(\lambda_k). \end{aligned}$$

Substituting  $\lambda_k$  and  $\beta_k$ ,  $k = 1, \dots, K-1$ , the last expression yields:

$$\begin{aligned} \prod_{k=1}^{K-1} B(\lambda_k, \beta_k) &= \frac{\Gamma\left(\sum_{k \in C_1} \alpha_{\tau_k^{-1}} + s_k(\boldsymbol{\xi}) + s_k(\mathbf{z})\right)}{\Gamma(\beta_0) \Gamma\left(\sum_{k \in C_1} \alpha_{\tau_k^{-1}} + s_k(\boldsymbol{\xi})\right)} \\ &\quad \times \prod_{k \in C_0} \Gamma(\alpha_{\tau_k^{-1}} + s_k(\boldsymbol{\xi}) + s_k(\mathbf{z})) \prod_{k \in C_1} \Gamma(\alpha_{\tau_k^{-1}} + s_k(\boldsymbol{\xi})). \end{aligned} \quad (\text{E.3})$$

Note here that  $\Gamma(\beta_0)$  does not depend on  $\boldsymbol{\xi}$  or  $\mathbf{z}$ , hence substituting Equations (E.2), (E.3) into (E.1) we obtain (17), as stated.

Next we proceed to deriving the distributions of  $\xi_i | \boldsymbol{\xi}_{[-i]}, \mathbf{z}, \mathbf{y}, c$  and  $z_j | \mathbf{z}_{[-j]}, \mathbf{z}, \mathbf{y}, c$ , for  $i = 1, \dots, r$ ;  $j = 1, \dots, s$ . Let us focus first at the probability a specific read  $i = 1, \dots, r$  of the first condition being assigned to a specific transcript  $k = 1, \dots, K$ , given the allocations of all remaining reads  $(\boldsymbol{\xi}_{[-i]}, \mathbf{z})$  and the state vector  $(c)$ . After discarding all irrelevant terms from Equation (17) we obtain that:

$$f(\xi_i | \boldsymbol{\xi}_{[-i]}, \mathbf{z}, c, \mathbf{x}) \propto \frac{\Gamma\left(\sum_{t \in C_1} \alpha_{\tau_t^{-1}} + s_t(\boldsymbol{\xi}) + s_t(\mathbf{z})\right)}{\Gamma\left(\sum_{t \in C_1} \alpha_{\tau_t^{-1}} + s_t(\boldsymbol{\xi})\right)} \prod_{t \in C_1} \Gamma(\alpha_{\tau_t^{-1}} + s_t(\boldsymbol{\xi}))$$

$$\times \prod_{t \in C_0} \Gamma(\alpha_{\tau_t^{-1}} + s_t(\boldsymbol{\xi}) + s_t(\mathbf{z})) f_{\xi_i}(x_i).$$

Now, notice that:  $s_t(\boldsymbol{\xi}) = s_t^{(i)}(\boldsymbol{\xi})$  for  $t \neq k$  while  $s_k(\boldsymbol{\xi}) = s_k^{(i)}(\boldsymbol{\xi}) + 1$  and recall that  $\Gamma(x+1) = x\Gamma(x)$ . Hence, the last equation simplifies to:

$$P(\xi_i = k | \boldsymbol{\xi}[-i], \mathbf{z}, c, \mathbf{x}) \propto \begin{cases} (\alpha_{\tau_k^{-1}} + s_k^{(i)}(\boldsymbol{\xi}) + s_k(\mathbf{z})) f_k(x_i), & k \in C_0 \\ \frac{\sum_{t \in C_1} \alpha_{\tau_t^{-1}} + s_t^{(i)}(\boldsymbol{\xi}) + s_t(\mathbf{z})}{\sum_{t \in C_1} \alpha_{\tau_t^{-1}} + s_t^{(i)}(\boldsymbol{\xi})} (\alpha_{\tau_k^{-1}} + s_k^{(i)}(\boldsymbol{\xi})) f_k(x_i), & k \in C_1 \end{cases}$$

which is Equation (18), as stated. Equation (19) is derived after following the similar arguments for  $z_j | \mathbf{z}_{[-j]}, \boldsymbol{\xi}, \mathbf{y}$ .

## F. Update of state vector in the rjMCMC sampler

In this section we introduce the reversible jump proposal for updating the state vector  $c$  and  $\mathbf{v}$ .

**Birth move:** Assume that the current state of the chain is

$$g := (c, \tau, \mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\xi}, \mathbf{z}).$$

We propose to obtain a new state for the chain

$$g = (c, \tau, \mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\xi}, \mathbf{z}) \rightarrow g' = (c', \tau', \mathbf{u}', \mathbf{v}', \boldsymbol{\theta}', \mathbf{w}', \boldsymbol{\xi}', \mathbf{z}'),$$

by a birth type move. This will increase the number of differentially expressed transcripts: either by one (if  $c_+ \geq 2$ ) or by two (if  $c_+ = 0$ ). At first, we choose a move of this specific type with probability proportional to the number of elements in  $C_0(c)$ , that is,  $K - c_+$ . Then, if  $c_+ \geq 2$  we select at random an index  $k_0 \in C_0(c)$  which we propose to add to  $C_1(c)$ . If  $c_+ = 0$  we select at random two indexes  $\{k_1, k_2\} \in C_0(c)$  which we propose to move to (the previously empty)  $C_1(c)$ . The probability of selecting such a move type is,

$$P_{\text{birth}}(c \rightarrow c') = \begin{cases} \frac{K-c_+}{K} \frac{1}{K-c_+} = \frac{1}{K}, & \text{if } 2 \leq c_+ \leq K-1 \\ \frac{K-0}{K} \frac{1}{\binom{K}{2}} = \frac{2}{K(K-1)}, & \text{if } c_+ = 0. \end{cases} \quad (\text{F.1})$$

Moreover, define the corresponding death probability

$$P_{\text{death}}(c \rightarrow c') = \begin{cases} \frac{c_+}{K} \frac{1}{c_+} = \frac{1}{K}, & \text{if } 3 \leq c_+ \leq K \\ \frac{2}{K}, & \text{if } c_+ = 2. \end{cases} \quad (\text{F.2})$$

If  $c_+ = 0$ , assume without loss of generality that  $k_1 < k_2$ . Then,  $C_1(c') = \{k_1, k_2\}$  and  $C_0(c') = \{1, \dots, K\} - C_1(c')$ . Now assume that  $c_+ \geq 2$ . It is obvious that in this case



$c'_k = c_k$  for all  $k \neq k_0$  and  $c'_{k_0} = 1$ . Moreover, the dead and alive subsets of the new state is obtained by deleting  $k_0$  from  $C_0(c)$  and adding it to  $C_1(c)$ . Let

$$j := \sum_{k \in C_1(c)} I(k < k_0) + 1 = \sum_{k=1}^{k_0} c_k + 1.$$

Then, the alive subset of the new state is

$$C_1(c') = \begin{cases} \{C_1(c)\}_k, & k < j \\ k_0, & k = j \\ \{C_1(c)\}_{k-1}, & j < k \leq c_+ + 1 \end{cases} \quad (\text{F.3})$$

and the dead subset will simply be  $C_0(c') = \{1, \dots, K\} - C_1(c')$ . Finally, the new permutation is defined as  $\tau' = (C_0(c'), C_1(c'))$ .

Now, we have to propose the values of  $\mathbf{u}'$ ,  $\mathbf{v}'$ . Recall that the dimension of  $\mathbf{u}$  is always constant, but the dimension of  $\mathbf{v}'$  will be increased by one. We consider them separately in order to keep it as simple as possible. For  $\mathbf{u}$  we propose to jump to a new state  $\mathbf{u}'$  which arises deterministically as the corresponding permutation of its previous values. In order to do this we just have to match the position of each element of  $\tau'$  in  $\tau$ . This means that

$$\mathbf{u}' = (\tau^{-1}\tau')\mathbf{u} = \tau'[(\tau^{-1}\mathbf{u})]. \quad (\text{F.4})$$

In order to construct a valid Metropolis-Hastings acceptance probability for the dimension changing move from  $\mathbf{v}$  to  $\mathbf{v}'$ , we should take into account the dimension matching assumption of Green (1995). In our set up, this assumption states that the jump from  $\mathbf{v} \rightarrow \mathbf{v}'$  should be done by producing one random variable that will bridge the dimensions, that is:

$$\mathbf{v}' = h(\mathbf{v}, \delta),$$

where  $\delta$  denotes a (univariate) random variable and  $h(\cdot, \cdot)$  an invertible transformation. We design this transformation following similar ideas from the standard birth and death moves of Richardson and Green (1997), Papastamoulis and Iliopoulos (2009). Let  $\delta \sim f_{\text{prop}}$ , where  $f_{\text{prop}}$  denotes the density function of a distribution with support  $(0, 1)$ . Then, the new parameter is obtained as

$$\mathbf{v}' = h(\mathbf{v}, \delta) := \begin{cases} (v_1(1-\delta), \dots, v_{j-1}(1-\delta), \delta, v_{j+1}(1-\delta), \dots, v_{c_+}(1-\delta)), & c_+ \geq 2 \\ (\delta, 1-\delta), & c_+ = 0 \end{cases} \quad (\text{F.5})$$

Finally, we have to compute the absolute value of the Jacobian of the transformation in (F.5). Now, recall that  $\mathbf{v}$  consists of  $c_+ - 1$  independent elements, so the dimension of the

Jacobian is  $c_+ \times c_+$  (and not  $(c_+ + 1) \times (c_+ + 1)$ ). Then, a routine calculation leads to:

$$|J(\delta, c)| = \begin{cases} (1 - \delta)^{c_+ - 1}, & c_+ \geq 2 \\ 1, & c_+ = 0 \end{cases} \quad (\text{F.6})$$

The new values of transcript expression  $\boldsymbol{\theta}', \mathbf{w}'$  are as follows. By Equations (5) and (F.4)

$$\boldsymbol{\theta}' = \tau'^{-1} \mathbf{u}' = \tau'^{-1} \tau'[(\tau^{-1} \mathbf{u})] \Rightarrow \boldsymbol{\theta}' = \boldsymbol{\theta}, \quad (\text{F.7})$$

and applying Equation (6):

$$\mathbf{w}' = \tau'^{-1} \left( \{u'_{\tau'_k} : k \in C_0(c')\}, \mathbf{v}' \sum_{k \in C_1(c')} u'_{\tau'_k} \right). \quad (\text{F.8})$$

Finally, we propose to reallocate all observations according to the new values  $\boldsymbol{\theta}'$  and  $\mathbf{w}'$ . This is simply done by using the full conditional distributions of  $\boldsymbol{\xi}', \mathbf{z}'$ . Let  $P(\boldsymbol{\xi}', \mathbf{z}' | \boldsymbol{\theta}', \mathbf{w}')$  denote the probability of the allocations, according to the general form given in Equations (13) and (14). Note here that such a reallocation it is not necessary, however it is suggested because improves the acceptance rate of the proposed move.

**SUPPLEMENTARY LEMMA 1.** *The acceptance probability of the birth move is  $\min\{1, A(g, \delta, g')\}$ , where*

$$\begin{aligned} A(g, \delta, g') = & \frac{f(\mathbf{x}, \mathbf{y}, \mathbf{z}', \boldsymbol{\xi}' | \boldsymbol{\theta}', \mathbf{w}') P(\boldsymbol{\xi}, \mathbf{z} | \boldsymbol{\theta}, \mathbf{w})}{f(\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\xi} | \boldsymbol{\theta}, \mathbf{w}) P(\boldsymbol{\xi}', \mathbf{z}' | \boldsymbol{\theta}', \mathbf{w}')} \\ & \times \frac{P_{\text{death}}(c' \rightarrow c) P(c') f(\mathbf{u}', \mathbf{v}' | \boldsymbol{\alpha}, \boldsymbol{\gamma}) |J(\delta, c)|}{P_{\text{birth}}(c \rightarrow c') f_{\text{prop}}(\delta) P(c) f(\mathbf{u}, \mathbf{v} | \boldsymbol{\alpha}, \boldsymbol{\gamma})}. \end{aligned} \quad (\text{F.9})$$

**PROOF.** See the acceptance probability in Green (1995).

Note that for the Jeffreys prior (2), (3) it holds that:

$$\frac{P(c')}{P(c)} = \begin{cases} \frac{\pi}{1-\pi}, & c_+ \geq 2 \\ \frac{\pi^2}{(1-\pi)^2}, & c_+ = 0. \end{cases}$$

**Death move:** A death proposal is the reverse move of a birth. Suppose that we propose a transition

$$g = (c, \tau, \mathbf{u}, \mathbf{v}, \boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\xi}, \mathbf{z}) \rightarrow g' = (c', \tau', \mathbf{u}', \mathbf{v}', \boldsymbol{\theta}', \mathbf{w}', \boldsymbol{\xi}', \mathbf{z}'),$$

via a death move. At first we choose at random an element of the alive subset and propose to move and paste it to the dead subset (in the case that the alive subset consists of only two transcripts, then we essentially setting the alive subset to the empty set). This reduces

the number of differentially expressed transcripts either by one (if  $c_+ \geq 2$ ) or by two (if  $c_+ = 2$ ).

If  $c_+ = 2$  let  $C_1(c) = \{k_1, k_2\}$  and  $\mathbf{v} = (v_1, v_2)$ , with  $v_2 = 1 - v_1$ . Then the random variable that we have to produce during the reverse move is deterministically set to  $\delta = v_1$ , and  $\mathbf{v}' = \emptyset$ . In any other case, assume that the chosen alive transcript index is  $k_1$ , then define

$$j := \sum_{k \in C_0(c)} I(k < k_1) + 1 = \sum_{k=1}^{k_1} c_k.$$

Then, the reverse transformation of (F.5) implies that

$$(\mathbf{v}', \delta) = h^{-1}(\mathbf{v}) := \begin{cases} \left\{ \left( \frac{v_1}{1-v_j}, \dots, \frac{v_{j-1}}{1-v_j}, \frac{v_{j+1}}{1-v_j}, \dots, \frac{v_{c_+}}{1-v_j} \right), v_j \right\}, & c_+ \geq 2 \\ v_j, & c_+ = 0 \end{cases} \quad (\text{F.10})$$

Everything works in a reverse way compared to the birth move, so the acceptance probability of a death move is then simply given by  $\min\{1, A^{-1}(g', v_j, g)\}$ .

## G. Update of state vector in the collapsed sampler

According to Equation (12), the conditional distribution of  $c$  is written as:

$$f(c|\boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto f(\boldsymbol{\xi}, \mathbf{z}|c, \mathbf{x}, \mathbf{y})f(c|\pi)h_c,$$

where  $f(c|\pi)$  denotes the prior distribution of  $c$  defined in Equation (3),  $f(\boldsymbol{\xi}, \mathbf{z}|c, \mathbf{x}, \mathbf{y})$  is defined in Equation (17) of Theorem 2 and  $h_c = \frac{\Gamma(\sum_{\ell=1}^{c_+} \gamma_\ell)}{\prod_{\ell=1}^{c_+} \Gamma(\gamma_\ell)}$  corresponds to the constant term of the prior distribution for  $\mathbf{v}$ . However, in order to fully update the state vector we would have to compute this quantity for all  $c \in \mathcal{C}$ , and this would be time consuming.

An alternative is to update two randomly selected indices, given the configuration of remaining ones. Hence, if  $j_1$  and  $j_2$  denote two distinct transcript indices, then we perform a Gibbs update to  $c_{j_1, j_2} | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}$ . Let  $d = \sum_{k \neq j_1, j_2} c_k$ . Since  $c_+ = \sum_k c_k \neq 1$ , we have to differentiate the subsequent procedure between the following cases:  $d = 0$ ,  $d = 1$  and  $d > 1$ . If  $d = 0$  then  $c_{j_1, j_2} \in \{(1, 1), (0, 0)\}$ . In case that  $d = 1$  then  $c_{j_1, j_2} \in \{(1, 1), (1, 0), (0, 1)\}$ . Finally, if  $d > 1$  then  $c_{j_1, j_2} \in \{(1, 1), (0, 0), (1, 0), (0, 1)\}$ . Hence, the following full conditional distribution is derived:

$$P(c_{j_1} = 1, c_{j_2} = 1 | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto f(\boldsymbol{\xi}, \mathbf{z}|c, \mathbf{x}, \mathbf{y})\pi^2 h_c \quad (\text{G.1})$$

$$P(c_{j_1} = 0, c_{j_2} = 0 | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto \begin{cases} f(\boldsymbol{\xi}, \mathbf{z}|c, \mathbf{x}, \mathbf{y})(1 - \pi)^2 h_c, & d \neq 1 \\ 0, & d = 1 \end{cases} \quad (\text{G.2})$$

$$P(c_{j_1} = 1, c_{j_2} = 0 | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto \begin{cases} f(\boldsymbol{\xi}, \mathbf{z} | c, \mathbf{x}, \mathbf{y}) \pi (1 - \pi) h_c, & d \neq 0 \\ 0, & d = 0 \end{cases} \quad (\text{G.3})$$

$$P(c_{j_1} = 0, c_{j_2} = 1 | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto \begin{cases} f(\boldsymbol{\xi}, \mathbf{z} | c, \mathbf{x}, \mathbf{y}) (1 - \pi) \pi h_c, & d \neq 0 \\ 0, & d = 0. \end{cases} \quad (\text{G.4})$$

SUPPLEMENTARY LEMMA 2. *The update of a randomly selected block of  $c$  in the collapsed sampler:*

(a) *Select randomly two distinct indices  $\{j_1, j_2\}$  from the set  $\{1, \dots, K\}$*

(b) *Update  $c_{j_1, j_2} | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}$  as detailed in Equations (G.1)–(G.4)*

*corresponds to a Metropolis-Hastings step in which the proposed state is always accepted.*

PROOF. Assume that the current state of the chain is  $g = (c, \boldsymbol{\xi}, \mathbf{z}, \pi)$  and we propose to move to state  $g' = (c', \boldsymbol{\xi}, \mathbf{z}, \pi)$ , where  $c'_k = c_k$  if  $k \neq j_1, j_2$  and  $c_{j_1, j_2}$  is drawn from the full conditional distribution. The proposal density in this case can be expressed as

$$P(g \rightarrow g') = \frac{1}{\binom{K}{2}} f(c'_{j_1, j_2} | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto \frac{1}{\binom{K}{2}} f(c', \boldsymbol{\xi}, \mathbf{z}, \pi | \mathbf{x}, \mathbf{y}) = \frac{1}{\binom{K}{2}} f(g' | \mathbf{x}, \mathbf{y}).$$

The probability of proposing the reverse move (from  $g'$  to  $g$ ) equals to

$$P(g' \rightarrow g) = \frac{1}{\binom{K}{2}} f(c_{j_1, j_2} | c_{-[j_1, j_2]}, \boldsymbol{\xi}, \mathbf{z}, \pi, \mathbf{x}, \mathbf{y}) \propto \frac{1}{\binom{K}{2}} f(c, \boldsymbol{\xi}, \mathbf{z}, \pi | \mathbf{x}, \mathbf{y}) = \frac{1}{\binom{K}{2}} f(g | \mathbf{x}, \mathbf{y}).$$

Thus, the Metropolis-Hastings ratio for the transition  $g \rightarrow g'$  is expressed as

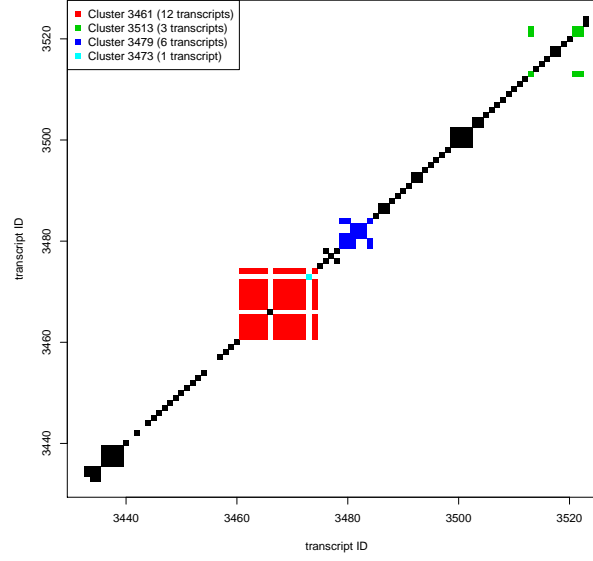
$$\frac{f(g' | \mathbf{x}, \mathbf{y}) P(g' \rightarrow g)}{f(g | \mathbf{x}, \mathbf{y}) P(g \rightarrow g')} = \frac{f(g' | \mathbf{x}, \mathbf{y}) \frac{1}{\binom{K}{2}} f(g | \mathbf{x}, \mathbf{y})}{f(g | \mathbf{x}, \mathbf{y}) \frac{1}{\binom{K}{2}} f(g' | \mathbf{x}, \mathbf{y})} = 1.$$

## H. Clustering of reads and transcripts

Let  $Q = (q)_{ij}$  be a  $K \times K$  symmetric matrix. For  $i = 1, 2, \dots, K$  and  $j = 1, \dots, K$  let  $N_{ij}$  denotes the number of reads that map to both transcripts  $i$  and  $j$ . Define:

$$q_{ij} := \begin{cases} 1 & \text{if } N_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,  $Q$  would be a diagonal matrix if all reads were uniquely mapped, but for real datasets  $Q$  is a sparse and almost diagonal matrix. A typical graphic representation of  $Q$  is illustrated in Figure 10 using a set of simulated reads from the *Drosophila Melanogaster* transcriptome. Each pixel corresponds to a pair of transcripts that contain at least one



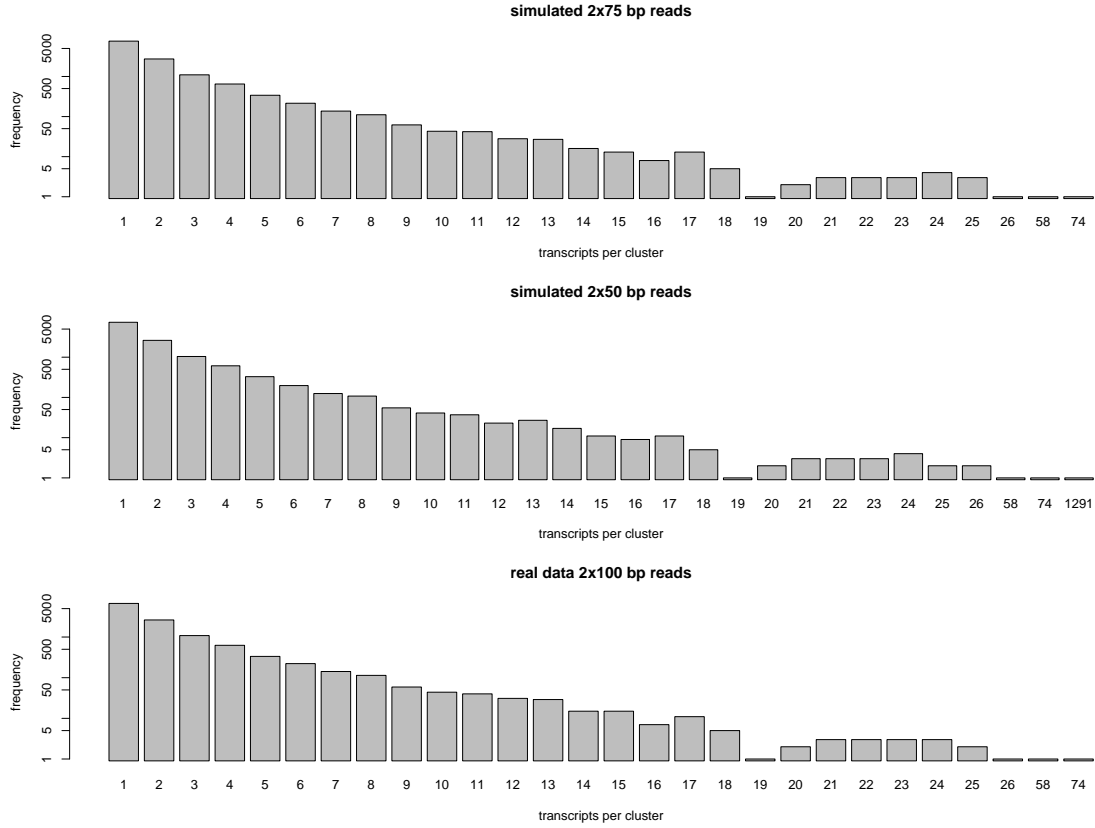
**Fig. 10.** Clusters of transcripts for a simulated set of 75 bp paired reads from the Drosophila transcriptome, containing  $K = 28763$  transcripts. For illustration purposes, only a subset consisting of 281 transcripts is shown and four clusters are emphasized using different colours. White colour corresponds to  $N_{ij} = 0$  aligned reads to both transcripts  $i, j$ .

read aligned to both transcripts of the pair. If all reads were uniquely aligned, this figure would consist only of the diagonal line and in this case each expressed transcript would form its own cluster. Note that the white gaps on the diagonal line indicate non-expressed transcripts. Many reads, however, map to more than one transcript resulting in clusters of transcripts, as the red, cyan, blue and green ones in Figure ???. The number of transcripts per cluster can have a wide range of values as displayed in Figure 11, but the majority of clusters consist of a very small number of transcripts compared to their total number. Next we formally define the notion of a cluster of transcripts.

**DEFINITION 3 (ASSOCIATED TRANSCRIPTS).** *Transcript  $k_1$  is associated to transcript  $k_2$  ( $k_1 \leftrightarrow k_2$ ) if  $q_{k_1 k_2} = 1$  or if exists a subset of indices  $\{i_1, i_2, \dots, i_m\} \subseteq \mathcal{K} := \{1, \dots, K\}$ ,  $m \geq 1$ , such that  $q_{k_1 i_1} + q_{i_1 i_2} + \dots + q_{i_{m-1} i_m} + q_{i_m k_2} = m + 1$ .*

**DEFINITION 4 (CLUSTER OF TRANSCRIPTS).** *The set of all associated transcripts of a given transcript  $k \in \mathcal{K} : \sum_{i=1}^K N_{ik} > 0$ :  $\mathcal{C}_k := \{j \in \mathcal{K} : j \leftrightarrow k\}$ , denotes the cluster of  $k$ .*

Note that according to definition 4:  $k_1 \leftrightarrow k_2 \Leftrightarrow \mathcal{C}_{k_1} = \mathcal{C}_{k_2}$ . We uniquely label each cluster by referring to its minimum index, as follows:



**Fig. 11.** Frequencies (in log-scale) of the number of transcripts per cluster using paired-end reads from the *Drosophila Melanogaster* transcriptome. Top and middle: (4489712) simulated reads, bottom: (22571142) reads from real data.

**DEFINITION 5 (CLUSTER LABELS).** *The label of cluster  $\mathcal{C}_k$  is defined as  $\mathcal{L}_\ell$ , with  $\ell = \min\{j \in \mathcal{C}_k\}$ . Conventionally, we set  $\mathcal{L}_0 := \{k \in \mathcal{K} : \mathcal{C}_k = \emptyset\}$ .*

Let  $n_c$  be the total number of clusters and assume that  $K_j$  is the number of transcripts associated with cluster  $\mathcal{L}_j$ . It holds that  $\cup_{j=1, \dots, n_c} \mathcal{L}_j = \mathcal{K}$  and  $\mathcal{L}_i \cap \mathcal{L}_j = \emptyset$  for  $i \neq j$ , that is,  $\{\mathcal{L}_0, \mathcal{L}_1, \dots, \mathcal{L}_{n_c}\}$  is a partition of  $\mathcal{K}$ . Finally, let  $r(\mathcal{L}_k)$  and  $s(\mathcal{L}_k)$  be the number of reads assigned to cluster  $\mathcal{L}_k$  from the first and second condition, respectively.

Next, assume that the proposed method is applied separately to each cluster. This would not lead to the same answer as the one with the full set of reads due to the fact that now each transcript weight corresponds to the relative expression inside each cluster. In order to ensure that the analysis will result to the same answer we should artificially augment each cluster with an extra pseudo-transcript that will contain information of the relative weight of each cluster. There are  $r(\mathcal{L}_j)$  and  $s(\mathcal{L}_j)$  reads from the first and second condition, respectively, exclusively aligned to cluster  $\mathcal{L}_j$ ,  $j = 1, \dots, n_c$ . Equivalently, there

are  $r - r(\mathcal{L}_j)$  and  $s - s(\mathcal{L}_j)$  reads from the first and second condition, exclusively aligning to the remaining clusters. Assume now that each cluster is augmented with an additional pseudo-transcript containing all remaining reads from both conditions. We conventionally set the label of the pseudo-transcript to  $K_j + 1$ . Given a set of reads from two biological conditions aligned to the reference transcriptome, the pipeline of the algorithm is the following.

- Partition the reference transcriptome and aligned reads into clusters.
- For each cluster  $j = 1, \dots, n_c$ , containing  $K_j \geq 1$  transcripts:
  - augment the cluster by the remaining pseudo-transcript containing  $r - r(\mathcal{L}_j)$  and  $s - s(\mathcal{L}_j)$  reads from the first and second condition, respectively.
  - Run the rjMCMC or the collapsed sampler.

The following Lemma ensures that it is valid to apply this sampling scheme per cluster in order to estimate the marginal posterior distribution of expression and differential expression for the set of transcripts assigned to each cluster. Apparently, this is not equivalent to simultaneously sampling from the joint posterior distribution of the whole transcriptome, which is computationally prohibitive, however the estimation of the marginal behaviour of each cluster is feasible and computationally efficient due to the dimension reduction.

**SUPPLEMENTARY LEMMA 3.** *Let  $\tilde{\theta}_j := (\{\theta_j; j \in \mathcal{L}_j\}, \sum_{k \neq \mathcal{L}_j} \theta_k)$ ,  $\tilde{w}_j := (\{w_j; j \in \mathcal{L}_j\}, \sum_{k \neq \mathcal{L}_j} w_k)$  denote the augmented transcript expressions for the first and second condition respectively and  $\tilde{c}_j := (\{c_j; j \in \mathcal{L}_j\}, c_{K_j+1})$ , at cluster  $j = 1, \dots, n_c$ . A priori assume:*

$$\tilde{\mathbf{u}}_j \sim \mathcal{D}_{K_j} \left( \{\alpha_j; j \in \mathcal{L}_j\}, \sum_{k \neq \mathcal{L}_j} \alpha_k \right) \quad (\text{H.1})$$

$$\tilde{\mathbf{v}}_j | \tilde{\mathbf{c}}_j \sim \mathcal{D}_{\tilde{c}_+} (\gamma_1, \dots, \gamma_{K_j+1}). \quad (\text{H.2})$$

*Then for each cluster  $j = 1, \dots, n_c$ , the parallel rjMCMC or collapsed algorithm converge to  $f(\tilde{\theta}_j, \tilde{w}_j, \tilde{c}_j | \mathbf{x}, \mathbf{y})$  and  $f(\tilde{\mathbf{c}}_j | \mathbf{x}, \mathbf{y})$ , respectively.*

**PROOF.** The distribution (H.1) is derived by (9) by applying the aggregation property of Dirichlet distribution, while distribution (H.2) is the same as (10) given  $\mathbf{c} = \tilde{\mathbf{c}}$ . Recall that according to Definition 4 there are  $\sum_{i=1}^r I(z_i = K_j + 1) = r - r(\mathcal{L}_j)$  and  $\sum_{i=1}^s I(\xi_i = K_j + 1) = s - s(\mathcal{L}_j)$  reads allocated to the component labelled as  $K_j + 1$  for cluster  $j$ . This means that the update scheme:

- (a) Update allocation variables  $\tilde{\xi}_j$  and  $\tilde{z}_j$  and set  $s_{K_j+1}(\tilde{\xi}_j) := r - r(\mathcal{L}_j)$ ,  $s_{K_j+1}(\tilde{z}_j) := s - s(\mathcal{L}_j)$ .
- (b) Update free parameters  $\tilde{u}_j$  and  $\tilde{v}_j$
- (c) Update expression parameters  $\tilde{\theta}_j$  and  $\tilde{w}_j$
- (d) Update state vector  $\tilde{c}_j$

updates the collapsed parameter vector:

$$\left( \{\theta_j; j \in \mathcal{L}_j\}, \sum_{k \neq \mathcal{L}_j} \theta_k \right), \left( \{w_j; j \in \mathcal{L}_j\}, \sum_{k \neq \mathcal{L}_j} w_k \right), (\{c_j; j \in \mathcal{L}_j\}, c_{K_j+1})$$

using the full conditional distributions for steps 1, 2, 3 and the reversible jump acceptance ratio in step 4 (in the case of rjMCMC sampler) or the random scan Gibbs step (in case of collapsed Gibbs). Hence it converges to  $f(\tilde{\theta}_j, \tilde{w}_j, \tilde{c}_j | \mathbf{x}, \mathbf{y})$ .

Note that the previous result assumes a fixed prior probability of DE. In practice, the prior probability of DE is a random variable, following the Jeffreys' prior distribution. Hence, the clustered sampling is equivalent to joint sampling only in case of fixed prior probability of DE. But we have found that this has not any impact in practice since according to our simulations the Jeffreys' prior outperforms the fixed prior probability of DE.

If the reads are sufficiently large, the clusters of transcripts are essentially genes (or groups of genes). It should be clear that the number of clusters as well as the cluster with the largest number of transcripts depends on the read length: if the read length is small, there will be many reads that map to multiple genes and in such a case all of these reads will form a very large cluster, as the one displayed in Figure 11 (middle) containing 1291 transcripts. The convergence of the MCMC algorithm for such clusters is questionable. However, even in such cases the majority of transcripts and reads are still forming a large number of small clusters. It is worth mentioning here that the large cluster is created by a very small number of reads: in total there are 417709 reads belonging to this cluster. However, the number of reads that actually map to more than ones genes is equal to 1474. Hence, we could break the bonds of this large number of transcripts by simply discarding or filtering out this small portion of reads.

## I. Initialization, burn-in and number of MCMC iterations per cluster

After partitioning the reads and transcripts into clusters, the rjMCMC or collapsed sampler is applied as previously discussed. For each run (MCMC per cluster),  $\text{mcmc}_n$  independent



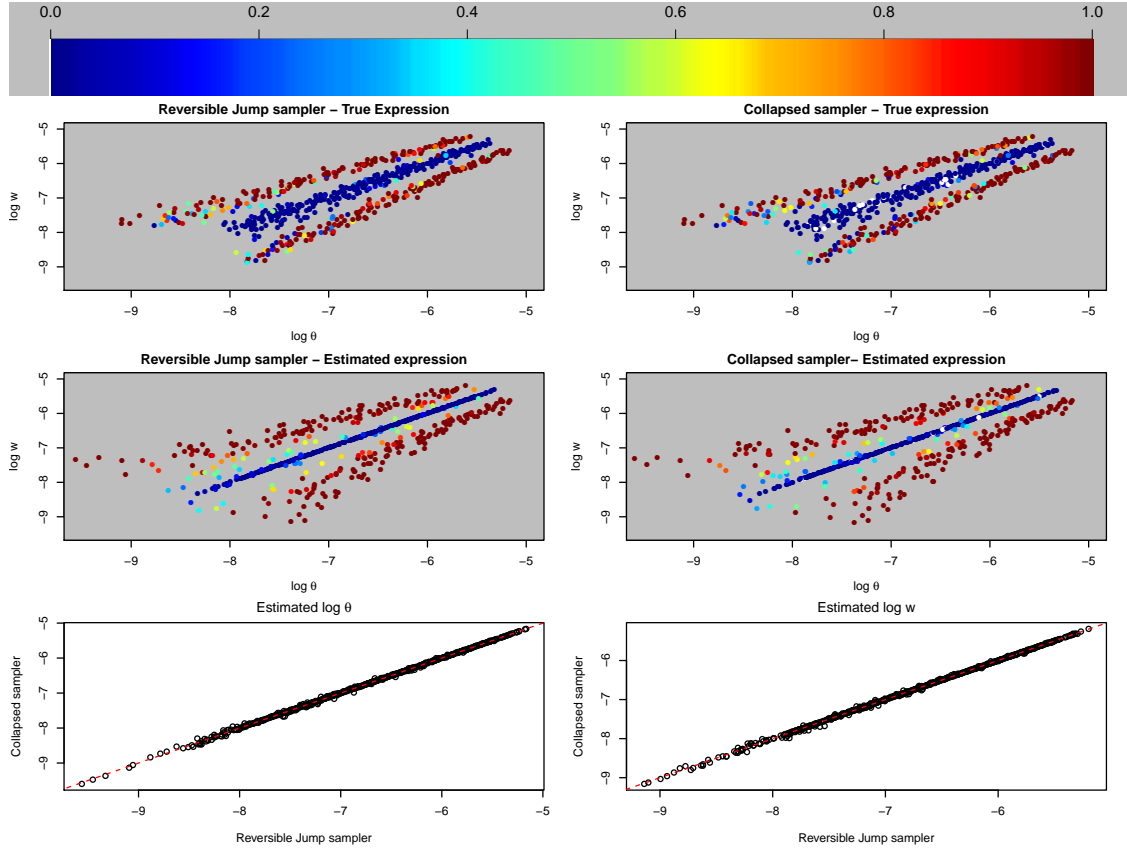
chains are obtained using randomly selected initial values for parameters  $\mathbf{u}$  and  $\mathbf{v}$ , drawn from (9) and (10). The first half of the chains is initialized from  $c_+ = 0$  (all transcripts are equally expressed), while the reverse (all transcripts are differentially expressed) holds for the initial state of the second half. The pseudo-transcript of each cluster (i.e. the mixture component labelled as  $K_j + 1$ ) is always initialized as differentially expressed. Given  $c, \mathbf{u}, \mathbf{v}$ , the initial relative transcript expressions are computed according to (7) and (8). Each chain runs for a fixed number of  $\text{mcmc}_m$  iterations, following a pre-specified number  $\text{mcmc}_b$  of burn-in draws. The posterior means are estimated by averaging the ergodic means across all chains, using a thinning of  $\text{mcmc}_t$  steps. The proposal distribution in the reversible jump step is an equally weighted finite mixture of Beta distributions:  $f_{\text{prop}} = \frac{1}{J} \sum_{j=1}^J \mathcal{B}(1, \beta_j)$ . All results reported are obtained using:  $\text{mcmc}_n = 6$ ,  $\text{mcmc}_m = 5000$ ,  $\text{mcmc}_b = 1000$ ,  $\text{mcmc}_t = 5$ ,  $J = 5$  and  $\{\beta_j; j = 1, \dots, 5\} = \{1, 10, 100, 250, 500\}$ .

## J. Comparison of samplers

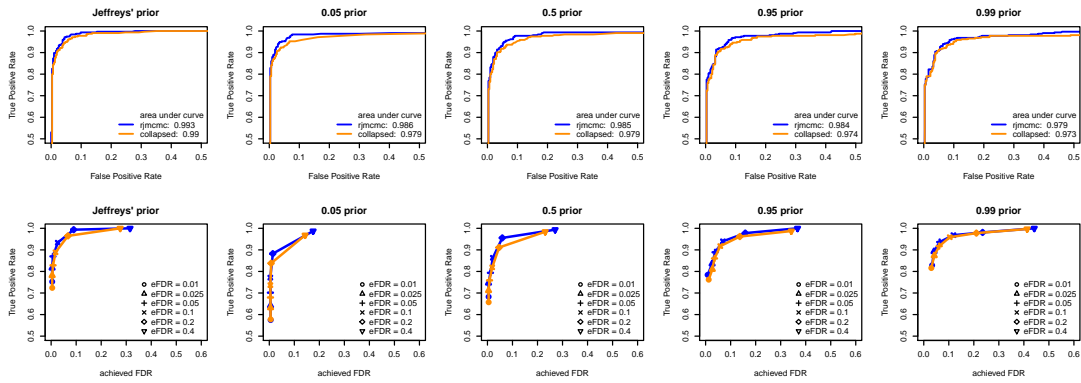
In this section we compare the Reversible Jump and the Collapsed version of our method as well we test the sensitivity of these samplers with respect to the prior probability of Differential Expression. In particular, we compare the Jeffreys' prior with a fixed prior probability of DE at 0.05, 0.50 and 0.95. We also examine the acceptance rates of the reversible jump proposal for updating the state vector  $c$ . Finally, a comparison between the clusterwise and raw sampler is made. For this purpose we used a toy example with relatively small number of reads and transcripts.

We simulated approximately 300000 reads per sample, which arise from a set of  $K = 630$  transcripts. The true values of the mixture weights used for the simulation are shown in Figure 12. Almost half of transcripts are Differentially Expressed and they correspond to the points that diverge from the identity line. The colour of each point corresponds to the posterior probability of differential expression for each sampler using the Jeffreys' prior distribution. The corresponding ROC curves for each sampler are shown in Figure 13, using also different prior distributions on the probability of differential expression. We conclude that the rjMCMC sampler tends to achieve higher true positive rate and a larger area under the curve. The achieved false discovery rates are shown at the second of 12. Compared to their expected values (eFDR) we conclude that both samplers achieve to control the False Discovery Rate at the desired levels, even when the prior favours DE transcripts (0.95 prior).

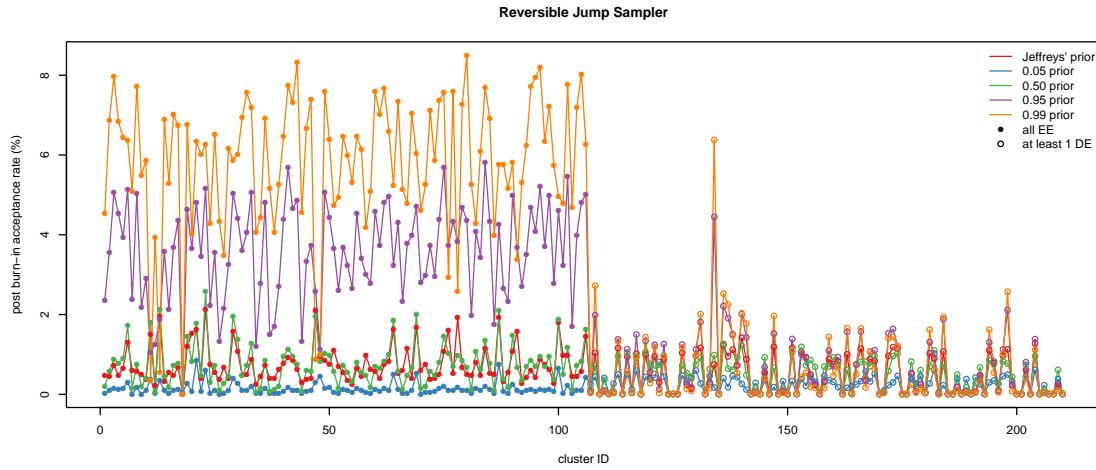
We also examine the acceptance rate of the reversible jump proposal, shown in Figure



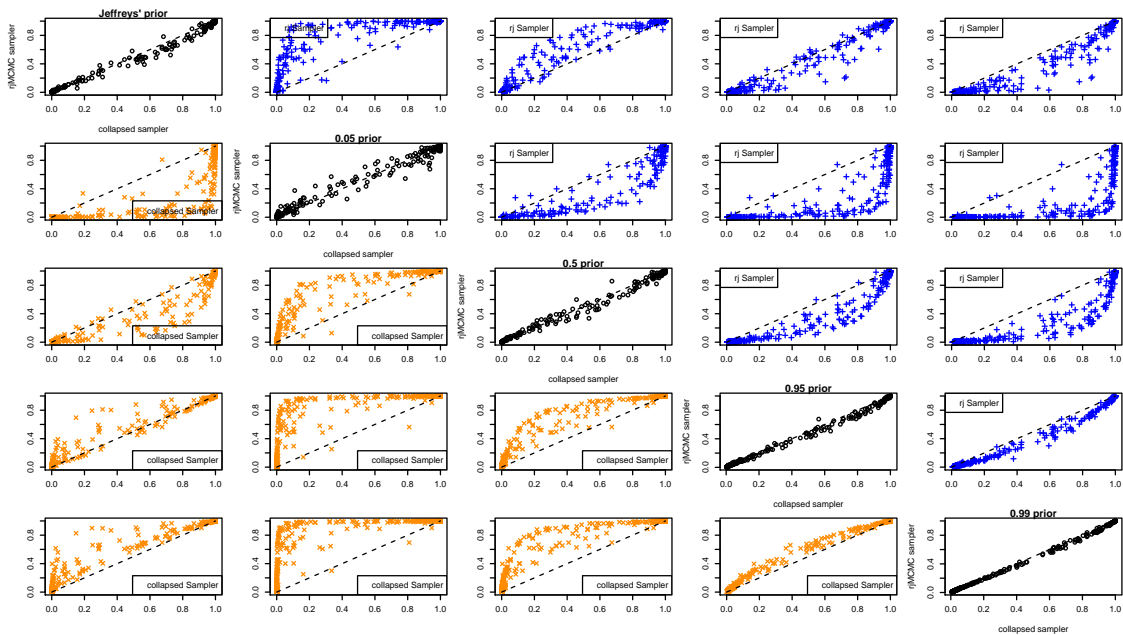
**Fig. 12.** True log-relative expression values for the toy example. The colour corresponds to the posterior probability of differential expression according to each sampler under the Jeffreys' prior (blue, green and red colors denote values close to 0, 0.5 and 1 respectively).



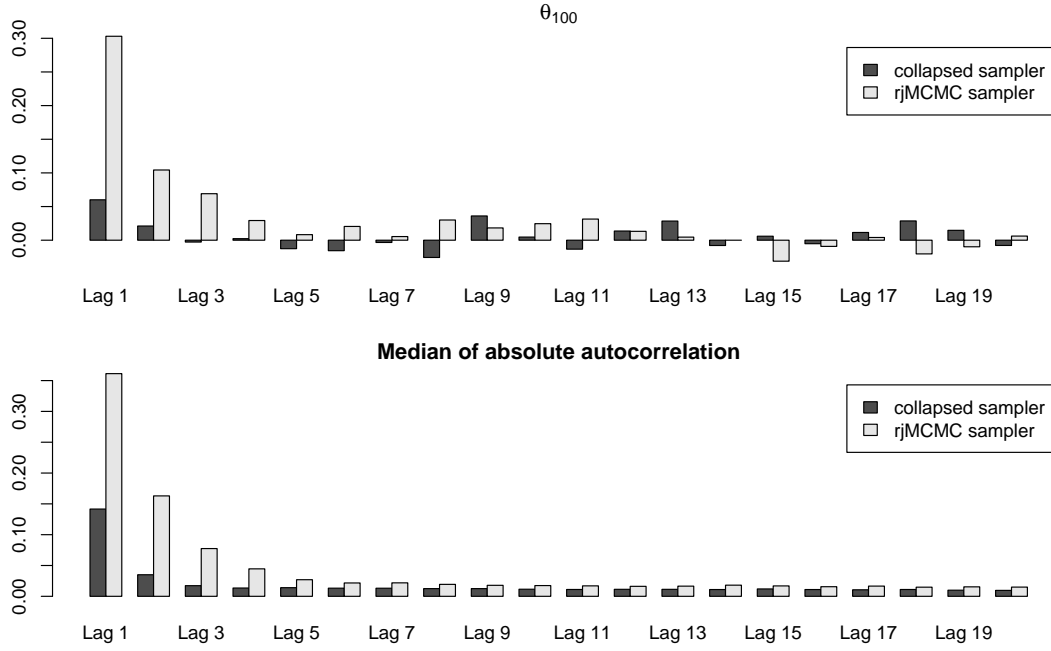
**Fig. 13.** ROC curves (up) and power-to-achieved FDR (down) for the toy example using different prior distribution on the probability of differential expression.



**Fig. 14.** Reversible Jump proposal acceptance rates per cluster (after discarding the MCMC draws which correspond to the burn-in period) for the update of  $c, v$  using different prior distributions. Note that the points are reordered so that clusters exclusively consisting of (truly) EE transcripts are shown first (solid points) followed by the clusters which contain at least one (truly) DE transcript (circles).



**Fig. 15.** Prior sensitivity of the rjMCMC (blue) and collapsed (orange) sampler. The main diagonal contains scatterplots of the posterior probability of DE between the rjMCMC and collapsed samplers for each prior distribution. The scatterplots of the same posterior probabilities for all possible prior combinations per sampler is shown at the upper (rjMCMC) and lower (collapsed) diagonal.

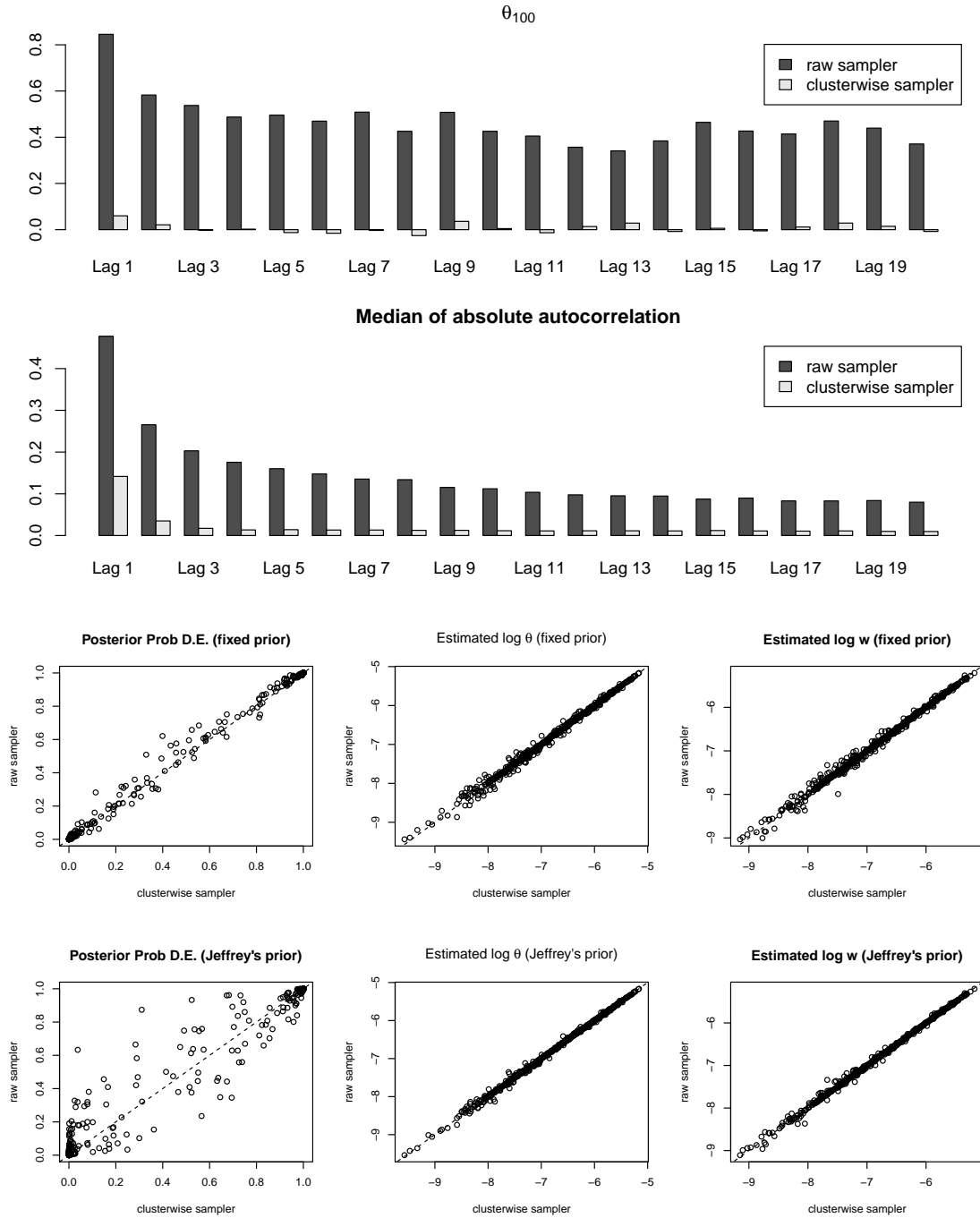


**Fig. 16.** Top: estimated autocorrelation function of the collapsed and rjMCMC algorithm for the sampled values of  $\log \theta_{100}$ . Bottom: Median of absolute autocorrelations for  $\log \theta_k$ ;  $k = 1, \dots, 630$ .

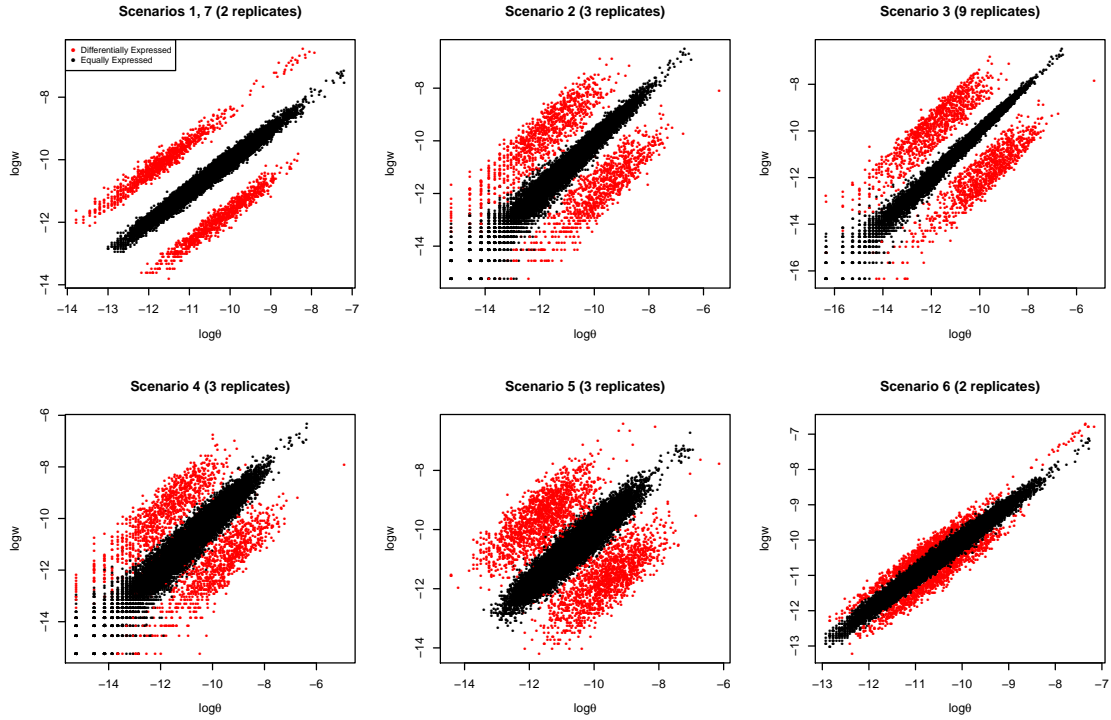
14. Overall, there is a small acceptance rate of proposed moves and there is a notable increase when the prior favours DE transcripts (0.95 or 0.99 prior). This mainly affects clusters consisting exclusively of EE transcripts. This conservative behaviour of rjMCMC sampler may indicate that the mixing of the algorithm is poor for the case of EE transcripts.

Next, we compare the autocorrelation function between the two samplers when using the Jeffrey's prior distribution for the probability of DE. A typical behaviour is shown in Figure 16 (top), displaying the autocorrelation function of  $\log \theta_k$  for a single transcript ( $k = 100$ ). In order to summarize the behaviour of autocorrelations across all  $K = 630$  transcripts we have computed the median of absolute autocorrelations for all  $\theta_k$ ;  $k = 1, \dots, K$ , as shown at the bottom of Figure 16. We conclude that the mixing of the collapsed sampler is notably better.

Finally we perform a comparison between the raw sampler (that is, taking into account the whole set of reads and transcripts) and the clusterwise one. For this reason we have run the raw collapsed MCMC sampler with a fixed prior of DE (equal to 0.5) as well as the Jeffrey's prior. As shown in Figure 17 (first two rows), the raw MCMC sampler exhibits very large autocorrelations compared to the clusterwise sampler (the autocorrelation function is nearly identical for both prior choices). The resulting estimates of DE and posterior



**Fig. 17.** Comparison of the clusterwise and raw MCMC algorithm. First row: estimated autocorrelation function for the sampled values of  $\log \theta_{100}$ . Second row: Median of absolute autocorrelations for  $\log \theta_k$ ;  $k = 1, \dots, 630$ . Third row: Comparison of estimated posterior means.



**Fig. 18.** Logarithm of true relative expression levels for seven simulation scenarios, averaged across the corresponding number of replicates.

means of transcript expression are shown in the third and fourth row of 17. Note that the estimates of posterior probability of DE exhibit larger variability under the Jeffrey’s prior. In both cases, the transcript expression estimates exhibit strong agreement. The number of iterations of the raw sampler was set to 2000000, following a burn-in period of 200000 iterations. Such a large number of iterations in general will not be sufficient in cases that the number of transcripts grows to typical values of RNA-seq datasets, hence running the raw MCMC sampler becomes prohibitive in general cases.

## K. Simulation study details

In the sequel,  $\mathcal{P}$  and  $\mathcal{NB}(\mu, \phi)$  denote the Poisson and Negative Binomial distributions respectively, where for the latter the parameterization with mean equal to  $\mu$  and variance equal to  $\mu + \mu^2/\phi$  is used,  $\mu \geq 0$ ,  $\phi > 0$ . Finally, let  $\text{RPK}_{jk}^{(A)}$  and  $\text{RPK}_{jk}^{(B)}$  denote the rpk values for transcript  $k$  at replicate  $j$  of condition A and B, respectively.

*Scenario 1 (2 Poisson replicates per condition)* Reads are simulated according to the following generative process.

$$\begin{aligned}
\mu_k &= 65, \quad k = 1, \dots, K, \quad n_d = 2g, \quad g = 1000 \\
\{k_1, \dots, k_{2g}\} &: \text{random sample of indices (without replacement)} \subseteq \{1, \dots, K\} \\
\delta_k^{(1)} &= 0.65, \quad k = k_1, \dots, k_g \\
\delta_k^{(2)} &= 3.25, \quad k = k_{g+1}, \dots, k_{2g} \\
(\mu_k^{(A)}, \mu_k^{(B)}) &= \begin{cases} (1, 1)\mu_k, & k \neq k_1, \dots, k_{n_d} \\ \left(\frac{1}{\delta_k^{(1)}}, \frac{1}{\delta_k^{(2)}}\right)\mu_k & k = k_1, \dots, k_g \\ \left(\frac{1}{\delta_k^{(2)}}, \frac{1}{\delta_k^{(1)}}\right)\mu_k, & k = k_{g+1}, \dots, k_{2g} \end{cases} \\
\text{RPK}_{jk}^{(A)} &\sim \mathcal{P}(\mu_k^{(A)}), \quad \text{RPK}_{jk}^{(B)} \sim \mathcal{P}(\mu_k^{(B)}), \quad k = 1, \dots, K, j = 1, 2.
\end{aligned}$$

The rpk values determined by this scenario used as input in Spanki and  $\approx 2400000$  reads per replicate are simulated ( $\approx 9600000$  reads in total). For non-differentially expressed transcripts, rpk values are simulated from a Poisson distribution with mean equal to 65 for both replicates of each condition. Next,  $n_d = 2000$  differentially expressed transcripts simulated with mean fold changes equal to  $\mu_k^{(A)}/\mu_k^{(B)} = 1/5$ ,  $k = 1, \dots, g$  and  $\mu_k^{(A)}/\mu_k^{(B)} = 5$ ,  $k = g + 1, \dots, 2g$ . More specifically, rpk values generated either from the  $\mathcal{P}(20)$  or  $\mathcal{P}(100)$  distribution. The averaged relative log-expression based on the true values are shown in Figure 18 and the points close to the identity line correspond to 26763 no-DE transcripts. The rest 2000 points that are far away from the identity line correspond to the DE transcripts. Apparently, this scenario corresponds to a clear cut case of separation between DE and non-DE transcripts at the two conditions.

*Scenario 2 (3 Negative Binomial replicates per condition)* Reads are simulated according to the following generative process.

$$\begin{aligned}
\mu_k &\sim \mathcal{U}(0, 70), \quad k = 1, \dots, K, \quad n_d = 2g, \quad g = 1000 \\
\{k_1, \dots, k_{2g}\} &: \text{random sample of indices (without replacement)} \subseteq \{1, \dots, K\} \\
\delta_k &\sim \mathcal{U}(\sqrt{3}, \sqrt{5}), \quad k = k_1, \dots, k_g \\
(\mu_k^{(A)}, \mu_k^{(B)}) &= \begin{cases} (1, 1)\mu_k, & k \neq k_1, \dots, k_{n_d} \\ (\delta_k, 1/\delta_k)\mu_k, & k = k_1, \dots, k_g \\ (1/\delta_k, \delta_k)\mu_k, & k = k_{g+1}, \dots, k_{2g} \end{cases} \\
\text{RPK}_{jk}^{(A)} &\sim \mathcal{NB}(\mu_k^{(A)}, 50), \quad \text{RPK}_{jk}^{(B)} \sim \mathcal{NB}(\mu_k^{(B)}, 50), \quad k = 1, \dots, K, j = 1, 2, 3.
\end{aligned}$$

The rpk values determined by this scenario used as input in Spanki and  $\approx 1335000$  reads per replicate are simulated ( $\approx 8010000$  reads in total). For non-differentially expressed transcripts, rpk values are simulated from the  $\mathcal{NB}(65, 50)$  for all three replicates of each condition. Next,  $n_d = 2000$  differentially expressed transcripts simulated with mean fold changes varying in the  $\mu_k^{(A)}/\mu_k^{(B)} \in (3, 5)$ ,  $k = 1, \dots, g$  and  $\mu_k^{(A)}/\mu_k^{(B)} \in (1/5, 1/3)$ ,  $k = g + 1, \dots, 2g$ . The averaged relative log-expression based on the true values are shown in Figure 18 and the points close to the identity line correspond to 26763 no-DE transcripts. The rest 2000 points correspond to the DE transcripts. Compared to Scenario 1, this case exhibits less separation between DE and non-DE transcripts at the two conditions due to (a) smaller fold changes, (b) increased replicate variability due to the Negative Binomial distribution and (c) larger range of transcript expression values.

*Scenario 3 (9 Negative Binomial replicates per condition)* The generative process is the same as Scenario 2 but with three times larger number of replicates per condition. In total  $\approx 24030000$  reads simulated. The averaged relative log-expression based on the true values are shown in Figure 18. Compared to Scenario 2, there should be more signal in the data in order to detect changes in expression due to the increased number of replicates.

*Scenario 4 (3 Negative Binomial replicates per condition, enhanced inter-replicate variance)* The generative process and the number of simulated reads is the same as Scenario 2 but with larger levels of variability among replicates. In particular we set  $\phi = 10$ , corresponding to five times larger variability compared to Scenario 2. The averaged relative log-expression based on the true values are shown in Figure 18. Compared to Scenario 2, there should be more uncertainty in the data in order to detect changes in expression due to the increased number of replicates.

*Scenario 5 (3 Negative Binomial replicates per condition, enhanced inter-replicate variance, smaller range for the mean)* The generative process and the number of simulated reads is the same as Scenario 4 but with more concentrated levels for the mean of true rpk values among replicates. In particular, we set

$$\begin{aligned} \mu_k &= 60, \quad k = 1, \dots, K, \quad n_d = 2g, \quad g = 1000 \\ \{k_1, \dots, k_{2g}\} &: \text{random sample of indices (without replacement)} \subseteq \{1, \dots, K\} \\ \delta_k &\sim \mathcal{U}(\sqrt{3}, \sqrt{5}), \quad k = k_1, \dots, k_g \end{aligned}$$



$$\begin{aligned}
(\mu_k^{(A)}, \mu_k^{(B)}) &= \begin{cases} (1, 1)\mu_k, & k \neq k_1, \dots, k_{n_d} \\ (\delta_k, 1/\delta_k)\mu_k, & k = k_1, \dots, k_g \\ (1/\delta_k, \delta_k)\mu_k, & k = k_{g+1}, \dots, k_{2g} \end{cases} \\
\text{RPK}_{jk}^{(A)} &\sim \mathcal{NB}(\mu_k^{(A)}, 10), \quad \text{RPK}_{jk}^{(B)} \sim \mathcal{NB}(\mu_k^{(B)}, 10), \quad k = 1, \dots, K, j = 1, 2, 3.
\end{aligned}$$

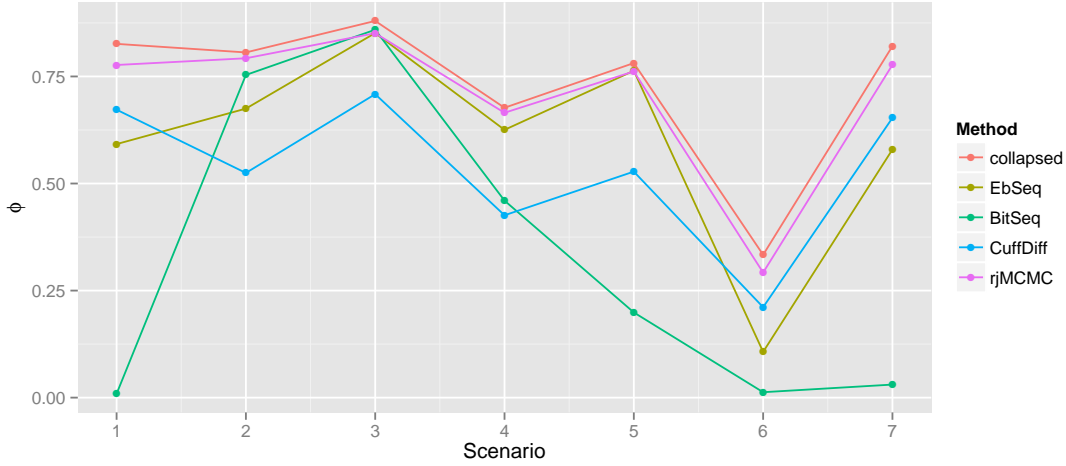
Note that the difference with Scenario 2 is that now  $\mu_k$  has a constant value for all  $k = 1, \dots, K$  and that the selected value of  $\phi$  results to five times higher dispersion. The averaged relative log-expression based on the true values are shown in Figure 18. It is obvious that now the range of relative expression values is smaller compared to Scenarios 2,3 and 4.

*Scenario 6 (2 Poisson replicates per condition, small fold change)* This is a revision of Scenario 1 under a smaller fold change between DE and EE transcripts. In this case we set  $\delta_k^{(1)} = 65/80$  and  $\delta_k^{(2)} = 65/50$ , resulting to a fold change of 1.6 for DE transcripts (instead of 5 as used at Scenario 1). As shown in Figure 18, the classification of DE and EE transcripts is not obvious.

*Scenario 7 (2 Poisson replicates per condition, unequal total number of reads)* This is a revised version of Scenario 1 under different sample sizes between the two conditions. Now the first condition contains approximately 46% larger amount of data than the second one. In particular, we simulated 2.81 and 1.93 million reads per replicate of the first and second condition, respectively. However, the relative expression levels are the same as in Scenario 1, as shown at the first plot of Figure 18.

Figure 19 displays the correlation between the true configuration of DE and EE transcripts and the estimated classification per method at the 0.05 level. Note that our collapsed sampler is ranked as the best method on every scenario. Moreover, our rjMCMC sampler is marginally the second best method. An interesting remark is that methods that control the false discovery rate exhibit a similar pattern across different scenarios, something that it is not the case for the standard BitSeq implementation. However note the improvement of standard BitSeq performance when the number of replicates is larger than two.

Figure 20 displays the ROC curves (left) and the true positive rate versus the achieved false discovery rate for the rjMCMC and collapsed samplers. The continuous lines correspond to the Jeffreys prior while the dashed lines correspond to a fixed probability of DE (equal to 0.5). The results are essentially the same for most scenarios. A notable



**Fig. 19.**  $\phi$ -coefficient between ground truth of DE and EE transcripts and the inferred classifications per method at the 0.05 level, for each simulation scenario.

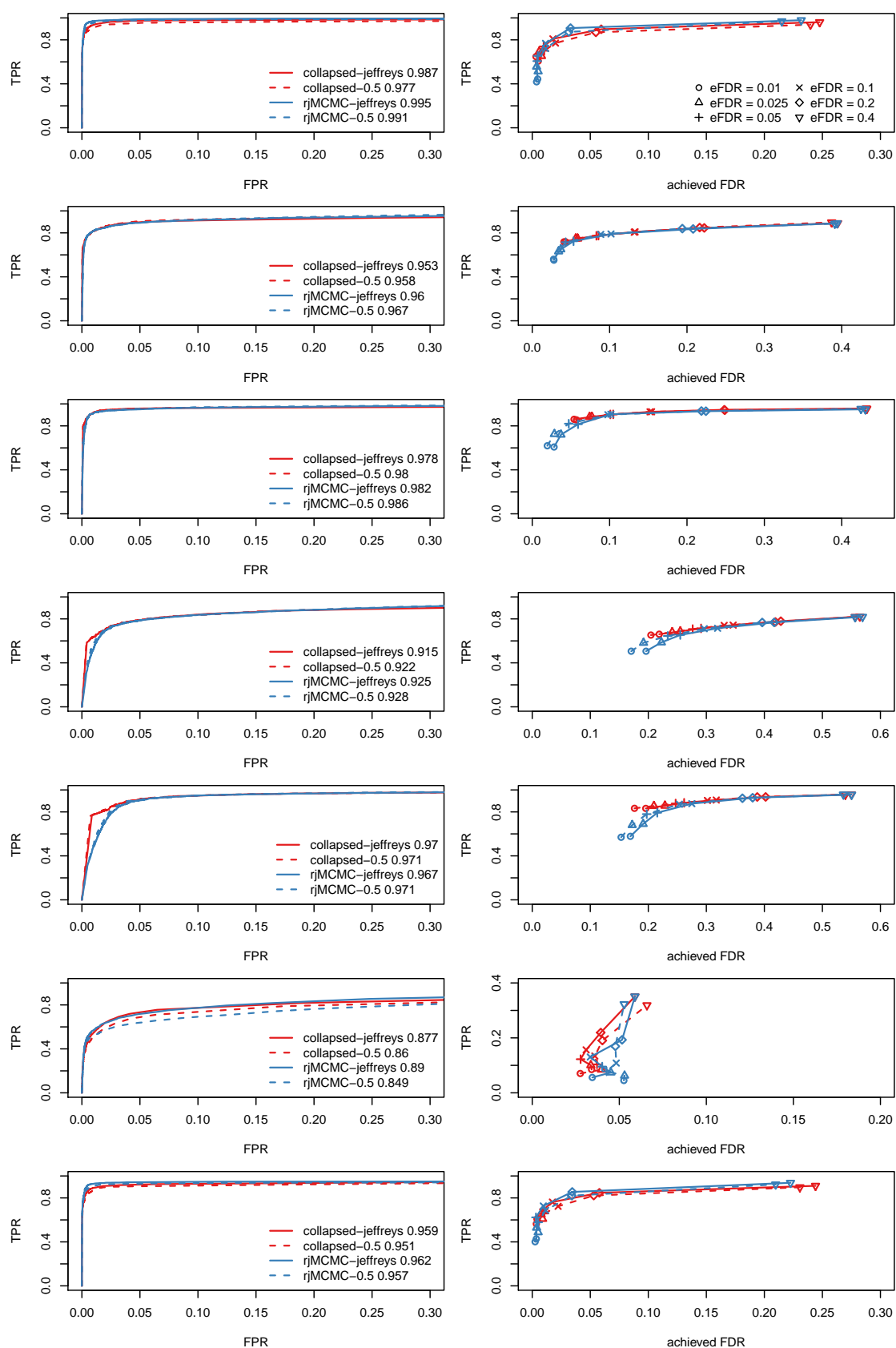
difference is observed at Scenario 6 where we conclude the superior performance of our method under the Jeffreys prior.

## L. Implementation of the algorithm

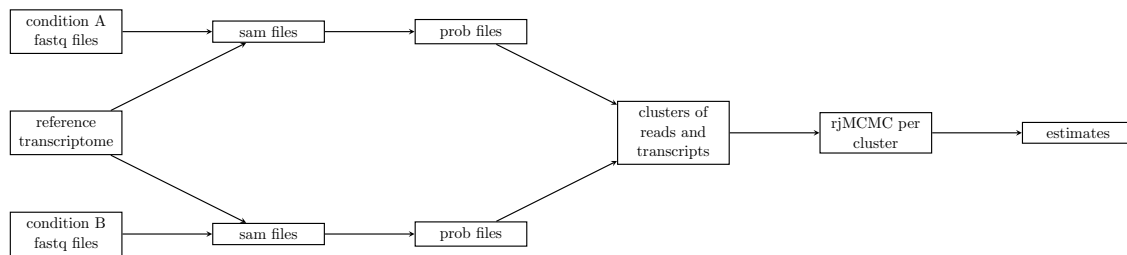
At first, the short reads (.fastq files) for each condition (A and B) are mapped to the reference transcriptome using Bowtie. The alignments (.sam files) are pre-processed using the `parseAlignment` command of BitSeq in order to compute the alignment probabilities for each read (.prob files). These files are used as the input of the proposed algorithm in order to (a) compute the clusters of reads and transcripts and (b) run the MCMC algorithm for each cluster. The output is a file containing the estimates of relative transcript expression for each condition and the posterior probability of differential expression.

Assume that there are two replicates per sample consisting of paired-end reads: `A1_1.fastq`, `A1_2.fastq`, `A2_1.fastq` and `A2_2.fastq` for sample A and `B1_1.fastq`, `B1_2.fastq`, `B2_1.fastq` and `B2_2.fastq` for sample B. Denote by `reference.fa` the fasta file with the transcriptome annotation. Let `outputRJ` and `outputCollapsed` denote the output directory of the rjMCMC and collapsed samplers, respectively. The following code describes a typical implementation of the whole pipeline, assuming that all input files are in the working directory (replace by the full paths otherwise).

```
# build bowtie2 indices and align reads
bowtie2-build -f reference.fa reference
bowtie2 -q -k 100 --no-mixed --no-discordant -x reference
```



**Fig. 20.** ROC curves (left) and power - to achieved plots (right) per simulation scenario, using different prior distribution on the probability of differential expression.

**Fig. 21.** General work-flow of the algorithm.

```

-1 A1_1.fastq -2 A1_2.fastq -S A1.sam
bowtie2 -q -k 100 --no-mixed --no-discordant -x reference
-1 A2_1.fastq -2 A2_2.fastq -S A2.sam
bowtie2 -q -k 100 --no-mixed --no-discordant -x reference
-1 B1_1.fastq -2 B1_2.fastq -S B1.sam
bowtie2 -q -k 100 --no-mixed --no-discordant -x reference
-1 B2_1.fastq -2 B2_2.fastq -S B2.sam

# compute alignment probabilities with BitSeq
parseAlignment A1.sam -o A1.prob --trSeqFile reference.fa
--uniform
parseAlignment A2.sam -o A2.prob --trSeqFile reference.fa
--uniform
parseAlignment B1.sam -o B1.prob --trSeqFile reference.fa
--uniform
parseAlignment B2.sam -o B2.prob --trSeqFile reference.fa
--uniform

# compute clusters and apply the rjMCMC sampler
rjBitSeq outputRJ A1.prob A2.prob C B1.prob B2.prob
# compute clusters and apply the collapsed sampler
cjBitSeq outputCollapsed A1.prob A2.prob C B1.prob B2.prob

```

The output of the rjMCMC and collapsed samplers is written to `outputRJ/estimates.txt` and `outputCollapsed/estimates.txt`, respectively. The overall work-flow is summarized in Figure 21.

**Table 1.**  $\phi$ -coefficient between the resulting classifications at the 0.05 level for HiSeq (lower diagonal) and MiSeq (upper) data.

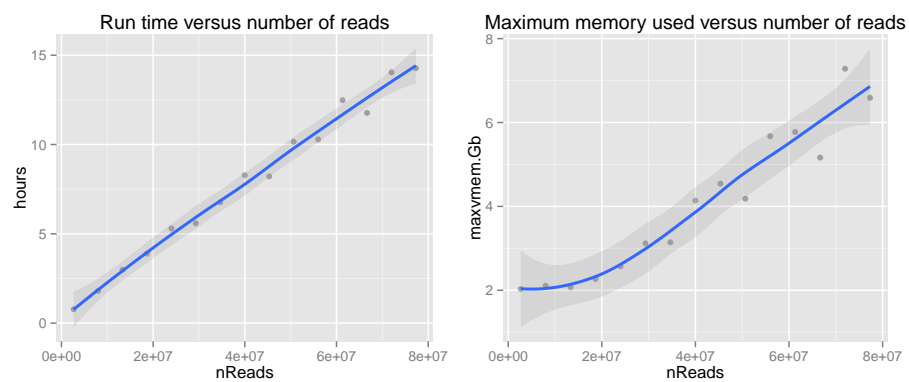
<i>Method</i>	<i>cuffdiff</i>	<i>BitSeq</i>	<i>EBSeq</i>	<i>cjBitSeq</i>
cuffdiff	1	0.43	0.32	0.32
BitSeq	0.64	1	0.58	0.59
EBSeq	0.52	0.61	1	0.70
cjBitSeq	0.56	0.63	0.75	1

**Table 2.** Approximate total number of reads (in millions) and run-time in hours for each example.

<i>dataset</i>	<i>reads</i>	<i>cufflinks</i>	<i>BitSeq</i>	<i>rsem/EBSeq</i>	<i>rjMCMC</i>	<i>collapsed</i>
scenario 1	9.4	0.9	4.4	2.2	4.8	2.7
scenario 2	8.0	0.8	3.3	1.8	4.5	3.4
scenario 3	24.0	2.1	9.1	6.2	9.8	8.4
scenario 4	8.0	0.9	3.4	2.5	4.4	3.2
scenario 5	14.0	1.1	9.7	3.7	6.6	5.1
scenario 6	9.4	0.8	4.5	1.9	4.3	2.6
scenario 7	9.5	0.7	5.8	1.7	4.1	2.5
MiSeq	21.3	1.0	4.8	2.4	6.8	3.9
HiSeq	97.0	2.4	22.3	11.1	26.1	19.8

## M. Additional tables and figures

Table 1 illustrates the correlation between the resulting classifications for the two real datasets in Section 3.3. Table 2 reports the running time needed for our experiments using 8 threads. The run-times reported for our method contains both cluster discovery and MCMC sampling. It should be mentioned that a significant portion of the reported run-times is allocated to the clustering part which is not optimized for speed (20% – 35% and 40% – 45% for the rjMCMC and collapsed samplers, respectively). More details regarding the computing time and memory usage demanded by our method are shown in Figure 22.



**Fig. 22.** Run time of the algorithm (left) and maximum virtual memory used (right) versus total number of (mapped) reads corresponding to the collapsed algorithm using 12 cores.